

**The Impact of the Item Types and Number of Solution Steps of Multiple-Choice Items
on Item Difficulty and Discrimination and Test Reliability**

By

Erkan Hasan Atalmis

Submitted to the graduate degree program in the Department of Psychology and Research in
Education and the Graduate Faculty of the University of Kansas in partial fulfillment of the
requirements for the degree of Doctor of Philosophy.

Chairperson Neal Kingston

Bruce Frey

Marianne Perie

Argun Saatcioglu

William Skorupski

Date Defended: 5/5/14

The Dissertation Committee for Erkan Hasan Atalmis certifies
that this is the approved version of the following dissertation:

The Impact of the Item Types and Number of Solution Steps of Multiple-Choice Items
on Item Difficulty and Discrimination and Test Reliability

Neal Kingston, Ph.D., Chairperson

Date approved: _____

Abstract

This study examined two multiple choice item-writing guidelines addressed by Haladyna, Downing, and Rodriguez (2002). One is related to using the “None of the Above (NOTA)” option, the other is about the plausible number of options for a multiple-choice item (MCI). These two guidelines were empirically tested using one-step and multi-step problems to identify their impact on item characteristics (item difficulty and item discrimination) and test characteristics (test reliability). Three forms with MCIs were generated and administered to approximately 1500 7th and 8th grade students in the United States and Turkey. Bi-factor Item Response Theory (IRT) was applied to assess dimensionality related to the number of solution steps of items. Multiple regression models were employed to determine the degree of impact these item-writing guidelines had on item and test characteristics for MCIs with one step solution (MCI with one-step solution) and those with more than one step solutions (MCIs with multi-step solution). The results show that item characteristics do not change significantly across the conventional MCIs with four options, MCIs with three options, and MCIs with NOTA option. The interaction between solution steps and the three MCI types had no significant impact on item characteristics. For the test with MCIs with a one-step solution, the findings demonstrate that four options are significantly more reliable than the NOTA options and not statistically different from three options. For the test with MCIs with multi-step solutions, four options are not statistically different from three and NOTA options. Compared to MCIs with four options, the results support that MCIs with NOTA options are preferable for MCIs with multi-step solutions while three options are desirable for both MCIs with one-step solutions and multi-step solutions.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Neal Kingston, for his insightful guidance to determine dissertation topic which fits my interests, experience background and academic abilities. His thoughtful comments and feedbacks always motivated me to make significant progress on my dissertation. This dissertation wouldn't have been completed without his valuable supports.

I also would like to express my appreciation to Dr. Argun Saatcioglu for encouraging me to choose the field of educational measurement and evaluation as my academic career.

Additionally, I would like to thank Dr. Bruce Frey, Dr. Marianne Perie, and Dr. William Skorupski for constructive comments and valuable suggestions. I am very honored to have them in my dissertation committee.

Finally, I would like to thank my wife, Hale Atalmis, my dearest and best friend in my life. There is no words to suitably express my feeling to her support. Her unwavering encouragement, support and love helped me have an efficient and pleasant dissertation process.

Table of Contents

Abstract.....	i
List of Tables	Error! Bookmark not defined.
List of Figures.....	vii
List of Equations	viii
CHAPTER 1	1
Introduction.....	1
Background	1
Statement of the Problem.....	5
Purpose of Study	6
Research Questions	7
Hypotheses.....	8
CHAPTER 2	9
Literature Review.....	9
Item-Writing Guidelines	9
None of the above.	11
Replacement method.....	12
Item statistics	13
<i>Item difficulty.</i>	13
<i>Item discrimination.</i>	15
Test statistics.....	17
<i>Test reliability.</i>	17
The number of options.	18
Elimination method.....	18
Item statistics.	19
<i>Item difficulty.</i>	19
<i>Item discrimination.</i>	20
Test characteristics.....	22
<i>Test reliability.</i>	22
<i>Test validity.</i>	23

Option Replacement and Elimination with Item Response function	23
Nonparametric smooth regression method.	24
The Number of Item Solution Steps with bi-factor IRT Model.....	26
Summary	27
CHAPTER 3.....	28
Methodology	28
Participants.....	28
Instrument Development.....	28
Procedure of item writing and test administration.	30
Item selection.	34
Option elimination with the Gauss Kernel smoothing regression method.....	35
Final Instrument	39
Data Collection.....	39
Determining number of steps for items.	40
Data Analysis	41
CHAPTER 4.....	43
Results.....	43
Confirming the Existence of Separate Factors Based on Number of Steps with Bi-Factor IRT Model.	43
Item Statistics.....	46
Testing Research Questions	48
Item Difficulty.	48
Item Discrimination.	49
Test Reliability.....	51
CHAPTER 5.....	55
Discussion	55
Summary of the Study	55
Implications for Testing Practice	59
Limitations and Directions for Future Research	59
References.....	62
Appendix A.....	68
Multiple – Choice Items.....	68
Appendix B	76

Option Analysis of Items	76
Appendix C	87
Factor Loading for Items in Bi-Factor IRT Model.....	87

List of Tables

Table 1. Three Formats of MCIs.....	11
Table 2. The relationship between the common core standards and the number of items	31
Table 3. The values of item discrimination and item difficulty for each item in pilot data	33
Table 4. The example of the items in different versions in the final test.....	35
Table 5. Types of method to eliminate the options.....	38
Table 6. Problems with one-step and multi-step solution.....	40
Table 7. Unidimensional and bi-factor model	44
Table 8. Summary of Item Characteristics	45
Table 9. Item Difficulty and Item Discrimination by Parallel Item Triplets	46
Table 10. Mean of Item Characteristic (Difficulty and Discrimination)	48
Table 11. Multiple Regression Model for Item Difficulty.....	48
Table 12. Multiple Regression Model for Item Discrimination	50
Table 13. Test Reliability for Different Set of the Test	53

List of Figures

Figure 1. An example of bi-factor model.....	26
Figure 2. Option Analysis of Item 21 in Form-A using Gauss Kernel Smoothing Regression Method.....	36
Figure 3. An example of the item in which content of options as elimination method is applied.	38
Figure 4. An example of unidimensional model.....	43
Figure 5. An example of bi-factor model.....	44

List of Equations

Equation 1. Distance Weight Equation	25
Equation 2. Gaussian kernel smoothing function	25
Equation 3. Gaussian kernel smoothing estimate function	25
Equation 4. The probability function of choosing option m for item j	25
Equation 5. The probability function of choosing option m for item j <i>with</i> latent variable.....	25
Equation 6. Multiple regression for item difficulty	48
Equation 7. Multiple regression for item discrimination	50
Equation 8. Standard error of estimate for reliability coefficient	51

CHAPTER 1

Introduction

Background

Multiple-choice items (MCI) are commonly used in standardized tests and classroom assessments for various disciplines in all fields (Haladyna, Downing, & Rodriguez, 2002; McCoubrie, 2004) due to their ability to obtain accurate and objective scores and to use time efficiently in administering a test and scoring it. For example, students can complete a test with a larger array of content with MCIs than when utilizing other item types, such as open-ended items (Collins, 2006). In other words, students complete tests with MCIs in less time and covering more information than with open-ended items when an equal number of items is included in the test. In addition, scoring a test with MCIs is completed more quickly than when scoring open-ended items. In particular, tests with MCIs can be scored rapidly via technology, such as bubble sheets or computer entry. Moreover, scoring is unbiased, objective, and accurate. Contrary to scoring open-ended items, different raters give the same scores for the same answer even when they are scored by hand since there is only one correct answer for a MCI. Therefore, a test with MCIs is widely applied in classroom and large-scale assessments with respect to time-efficiency of test administration and rating, and accuracy and objectivity of test scoring.

In technical terms, MCIs are commonly used to construct a reliable and valid test in order to measure accurately test takers' skills and ability. Reliability is defined as the consistency of a test result over time or with alternative forms, though it is often used based on a single test administration.

Different methods are employed to measure test reliability, such as test-retest, parallel-forms, and internal consistency. The same test is administered to the same group of people at two

different times in the test-retest method while in the parallel-forms method two parallel forms consisting of the items with similarity of construction and content are administered to one group of people in the parallel-forms method. In other words, two administrations for one single form are required in the former method whereas one single administration for two forms is required in the latter. Therefore, the degree of correlation between the forms from two administrations presents test-retest reliability in the test-retest method whereas the degree of correlation between two parallel forms indicates parallel-forms reliability.

Contrary to test-retest and parallel forms methods, the internal consistency method is designed as one single administration which saves time and money. With the internal consistency method, the items of a test are randomly split into two sets of groups, and the correlation between the items from the two groups are calculated to find split-half reliability. The average of all possible split-half reliability points out Cronbach's Alpha (α), which is the lower bound of the test reliability. Kuder-Richardson Formula 20 (KR-20) is utilized when dichotomous data is applied (0-1 or wrong-right response). In other words, α for continuous choices is used, and *KR-20* for nominal choices. In conclusion, the internal consistency method is the most widely used to interpret test reliability due to its more practical application than other methods in terms of administration and time.

Consequently, test designers interpret a test result by acknowledging a reliability coefficient for that test. For example, designers predict each examinee's true score by using the test reliability coefficient, which means that when the examinee takes the same test many times, the average of the scores he/she receives on each occasion indicates his/her true score. A true score is more accurately predicted with a reliable test. Moreover, when the test is reliable, the degree of the test validity can be investigated. In other words, it is a psychometrical fact that a

test which is not reliable is also not valid. Therefore, test reliability is the first important step to concluding that a test is valid.

Validity, which is related to score interpretation, is a commonly used term in the testing field; however, it is difficult to define it in less technical and simple words. Basically, validity is how well a test measures what it intends to measure. In terms of unitary concept, test validity is identified based on multiple sources of the evidence, such as evidence based on test content, evidence based on response processes, evidence based on internal structure, and evidence based on relations to other variables (AERA/APA/NCME Standards, 1999). Evidence based on test content is related to the content of a test, and its measurement includes some important steps. First, what is measured is the sample of knowledge in a particular course or program, which is defined as domain. A test can consist of more than one domain. Second, what students should know is determined, which is defined as an achievement standard. More than one standard is usually included in a single domain. Third, the item type - such as MCI and open-ended items, as well as the number of items – is determined and written for each standard with regard to their measurement priority. To illustrate evidence based on test content for validity evidence, assume a 7th grade test with 20 MCIs is designed. A domain of “equation and expression” is selected from the Common Core State Standards (2010), which is the real-life application based approach that will be applied in most states in the U.S. The domain includes four standards, one of which states: “Apply properties of operations as strategies to add, subtract, factor, and expand linear expressions with rational coefficients,” (CCSSI, 2010, p.49) and then five MCIs are written for each standard to design alignment study for evidence based on test content. Therefore, evidence based on test content is called a framework of test design.

Evidence based on response process is the evidence based on test takers' responses (AERA/APA/NCME Standards, 1999). The scores of test takers plays an important role in interpreting the test. For example, the scores can be classified as a label of a particular psychometrical ability (construct) definition for the test. Therefore, empirical studies should include appropriate data collection and evaluation method for construct definition of a test (AERA/APA/NCME Standards, 1999). In addition, the test-content has an important role in determining the definition of the construct. To illustrate with the example for evidence based on test content given above, a test with math items is taken by test takers. Through test takers' responses of "equation and expression" items and appropriate data analyzing method, the construct definition is identified.

The other evidence source of validity is the evidence based on relationships to other variables, empirical evidence comparing the test scores with other criterion to infer how well the test results perform (AERA/APA/NCME Standards, 1999). To illustrate with the "equation and expression" math test above, it is investigated how well and perfectly the test shows students' performance. Students' math performance from the previous year as criterion is employed to calculate convergent validity, which is the correlation between students' math performance from a previous year and the "equation and expression" test in this example. Conversely, when students' reading performance as criterion is used, the relationship between students' reading performance and the "equation and expression" is expected to be less correlated. This shows divergent validity. Therefore, convergent and divergent validity give some evidence about the degree of test validity; however, it is not enough to conclude that a test is valid either.

Consequently, validity plays an important role in improving and drawing conclusions from tests (AERA, APA, & NCME, 1999). A test lacking validity is not a good predictor for

decision-making. For instance, when an invalid classroom math test is applied, the teacher will make incorrect decisions about students' math ability. More particularly, while some students may not understand a particular item because the item is not written clearly, some cannot solve a particular question because it was not taught before. Therefore, only a valid test allows the test designers to properly interpret results and to fairly make decisions.

Statement of the Problem

Studies in the past have shown that writing item choices are a difficult part of the item-writing process (Rich & Johanson, 1990; Haladyna & Downing, 1989a; Hansen, 1997). Particularly, finding plausible distractors (guideline #29) is the most crucial part of writing a MCI. When a conventional MCI with four-options is written, three of the options should be distractors while one should be the key. Distractors should be written to reflect students' common errors to make a valid MCI (Haladyna & Downing, 1989a; Haladyna & Downing, 1989b; Haladyna et al., 2002). Thus, constructing a valid MCI is time-consuming, which is a limitation of using MCIs (Hansen, 1997; Burton et al., 1990). Writing distractors with students' common errors is the most challenging part of item-writing process.

There are different ways to write a MCI with fewer numbers of distractors. For example, one is to use "None of the Above," or NOTA, as an option because NOTA options could work better than a weak distractor or ineffective distractor (Rich & Johanson, 1990); however, NOTA options are not commonly used in classroom and large-scale assessment. Haladyna et al. (2002) state, as #25 of the writing-item guidelines: "Use carefully *None of the above*" (p. 314). This is because empirical studies indicated conversional results in terms of item characteristics (item difficulty, defined as the proportion of the students choosing correct answers and item

discrimination, defined as how well the item discriminates between students with high ability and low ability) and test characteristics (test reliability and test validity-test criterion validity). However, there is not enough empirical evidence to definitely make a case for using the NOTA-option; more research is needed.

Other ways to decrease the work required to develop good distractors is to decrease the number of options, which is related to #18 of the writing-item guidelines that Haladyna et al. (2002) addressed, for example, writing a MCI with three options rather than with four-options. Although the most empirical studies related to this item guideline are applied (Haladyna et al., 2002; Frey et al., 2005), the findings of the studies are contradictory about the impact of the number of options on item characteristics (item difficulty and item discrimination) and test characteristics (test reliability and test validity-test criterion validity) over the past 25 years. Likewise, similar to the guideline for using the NOTA-option (#25), more research is necessary due to controversial results from the previous studies.

Purpose of Study

Constructing MCIs with fewer numbers of options or with the NOTA-option was advocated by several researchers (Crehan, Haladyna, & Brewer, 1993; Knowles & Welch, 1992); however, only a limited number of empirical studies were employed in the past. These empirical studies interpreted the quality of a MCI with fewer numbers of options or with the NOTA-option in terms of item characteristics and test characteristics. In other words, the impact of the number of options or NOTA-option on item and test characteristics was investigated in these studies; however, no study has determined how the impact of the number of options or NOTA-option on item and test characteristics could change for MCIs with a one-step solution and multi-step

solutions. This study will respond to these issues by administering a math test with 30 MCIs to approximately 1650 7th and 8th grade students from the U.S. and from Turkey. A single test with three forms constructed with parallel questions is used in this study for each country. The forms are designed in the test respectively: 10 MCIs with four options, 10 MCIs with three options, and 10 MCIs with NOTA among the four options. The single test is administered in one class period of between 40 and 50 minutes. Multiple-Regression Analysis and t-test are conducted. The dependent variables are item difficulty, item discrimination, and test reliability coefficient. Independent variables are item types (MCIs with four options, MCIs with three options, and MCIs with NOTA-option), number of solution steps of the item (one-step solution and multi-step solutions).

Research Questions

This study addresses four research questions:

1. Do item characteristics (item difficulty and item discrimination) change when NOTA as an option is applied for MCIs with a one-step solution and multi-step solutions?
2. Does test reliability change when NOTA as an option is applied for MCIs with a one-step solution and multi-step solutions?
3. Do item characteristics (item difficulty and item discrimination) change when the number of options decreases from 4 to 3 for MCIs with a one-step solution and multi-step solutions?
4. Does test reliability change when the number of option decreases from 4 to 3 for MCIs with a one-step solution and multi-step solutions?

Hypotheses

The following hypotheses are examined in the current study on the basis of the information about these two guidelines:

1. Item difficulty (p) should decrease statistically when NOTA as an option is applied for MCIs with a one-step solution and multi-step solutions.
2. Item discrimination (r) should not change significantly when NOTA as an option is applied for MCIs with a one-step solution and multi-step solutions.
3. Test reliability should not change significantly when NOTA as an option is applied for MCIs with a one-step solution and multi-step solutions.
4. Item difficulty (p) should not change significantly when the number of options decreases from 4 to 3 for MCIs with a one-step solution and multi-step solutions.
5. Item discrimination (r) should not change significantly when the number of options decreases from 4 to 3 for MCIs with one-step solution and multi-step solutions.
6. Test reliability should not change significantly when the number of options decreases from 4 to 3 for MCIs with one-step solution and multi-step solutions.

CHAPTER 2

Literature Review

This chapter begins by delineating the item-writing guidelines used for constructing multiple-choice items (MCI) for reliable and valid standardized tests of achievement, both large-assessment and classroom evaluation. Writing MCIs without considering these guidelines can make the items weak and ineffective, which affects item characteristics (item difficulty and discrimination) and test characteristics (test reliability and validity). These four characteristics, which will be discussed in this section, are the keystones for researchers and item writers to identify how the item-writing guidelines affect item and test quality. Therefore, item-writing guidelines could play an important role in constructing high-quality items which are fundamental elements for an effective and objective test.

We also illustrate nonparametric smooth regression method to determine option analysis for MCIs. In the last section, we show how to determine number of solution steps of MCI by bi-factor IRT model.

Item-Writing Guidelines

Item-writing guidelines are those used to write high-quality MCIs with respect to structure, content, and formatting (Haladyna, Downing, & Rodriguez, 2002). Item-writing guidelines have been addressed by many researchers in the past (Case & Swanson, 1998; Haladyna & Downing, 1989a; Haladyna & Downing, 1989b; Haladyna, Downing, & Rodriguez, 2002), however, Haladyna et al. (2002) made a unique study, broadly and completely conducting a literature review of empirical and non-empirical studies of the item-guidelines since 1990.

They identified 31 valid item-writing guidelines and classified them in five categories: content, formatting, style, forming the stem, and forming the choices. Nearly half the item-writing guidelines, 14 out of 31, were found to be related to forming choices including key (correct answer) and distractors (incorrect answers). Haladyna et al. found that writing item options are the primary part of writing an item. Therefore, item quality is affected by its option's quality in terms of content, structure, and formatting.

Studies in the past have shown that writing item options is a challenging part of the item writing process (Rich & Johanson, 1990; Haladyna & Downing, 1989a; Hansen, 1997) because the options include one key and distractors, more importantly, plausible distractors (those plausible errors the students widely make). Haladyna et al. (2002) emphasized this in guidelines #29, and #30, “make distractors plausible” and “use common error of students,” respectively; this means that for item-writers to generate plausible distractors they need content knowledge and time, whereas to find the common errors of students requires teaching experience. Therefore, generating reasonable distractors depends on other factors, which makes writing options difficult for item writers.

Some alternatives are used to make writing options easy and straightforward in terms of item effectiveness. For instance, the fewer number of options or “None of the Above (NOTA)” as a last option could be employed to construct a MCI. Haladyna et al. (2002) discussed these alternative methods via other guidelines, such as #18, “Write as many plausible distractors as you can” and #25, “Use carefully *None of the above*” (p. 341). Table 1 illustrates a MCI with four options, three options, and a NOTA-option. In other words, a MCI is given in three formats, but with the same content.

Table 1. Three Formats of MCIs

1. Conventional MCI with four-options	2. Conventional MCI with three-options	3. NOTA-item
What is $\frac{1}{3} + \frac{3}{4}$?	What is $\frac{1}{3} + \frac{3}{4}$?	What is $\frac{1}{3} + \frac{3}{4}$?
a. $\frac{13}{12}$ (Key)	a. $\frac{13}{12}$ (Key)	A. $\frac{13}{12}$ (Key)
b. $\frac{4}{7}$	b. $\frac{4}{7}$	B. $\frac{4}{7}$
c. $\frac{13}{24}$	c. $\frac{13}{24}$	C. $\frac{13}{24}$
d. $\frac{13}{34}$		D. None of the Above

These two guidelines, though, are not commonly used in large-scale and classroom test creation because there is not sufficient empirical research to support these guidelines. Recently, Frey and his colleagues (2005) analyzed the 20 best known assessment text books. Thirty percent of the books cited the number of options (guideline #18) while 75% cited NOTA-items (guideline #25). Many studies stated whether or not guidelines #18 and #25 are applied in constructing a MCI is controversial, debatable results have been found, and the empirical studies are limited. Therefore, these two guidelines are empirically tested in this study.

None of the above.

The number of empirical studies about NOTA-option is very limited over the past 25 years with only seven empirical studies supporting the NOTA-option, but in different ways. Some of the studies investigated the impact of the NOTA-option on item characteristics (item

difficulty and item discrimination), while others focused on test characteristics (test reliability). In other words, item and test characteristics were tested to determine how applicable NOTA-options are for MCI creation. Therefore, the acceptability of NOTA-items to be employed in the standard test of the large-assessment and classroom evaluation is going to be discussed.

Replacement method.

Replacement method is a type of method for embedding NOTA as an option for a MCI. Namely, this method is applied by choosing a response-option, which is replaced by NOTA. For example, the third item in Table 1 has the NOTA-option, where the last response-option of the first item is replaced by NOTA. The last response-option of the first item is randomly chosen and replaced by a NOTA-option in this example. Other response-options could be chosen and substituted with the NOTA-option, but, how this replacement takes place can be performed by applying different methods. Thus, various possible replacement methods are plausible for the NOTA-option.

Although different methods for embedding NOTA as a response-option in a MCI were explored in studies over the past 25 years, four out of seven of the empirical studies explicitly provided replacement methods they used. For instance, some studies recommended that the NOTA-option should replace a weak distractor; some suggested that the NOTA-option should replace a strong distractor or a distractor randomly selected; and a few suggested that NOTA should be added as an alternative option (Wesman & Bennet, 1946; Hughes & Trimble, 1965; Dudycha & Carpenter, 1973). However, no research empirically identified which degrees of weakness and strength were applied with each method.

Frary (1991) applied a method for selecting a distractor randomly and replacing it by NOTA. Although other methods require two administrations, the pilot and final administration, this is the only method which requires just one administration, saving time and money. However, the best distractors may be unintentionally replaced by NOTA in this method, resulting in biased evaluation (Frary, 1991). Two studies replaced the least frequently chosen distractor with the NOTA-option (Rich & Johanson, 1990; Crehan et al., 1993). In this method, the percentage of students that selected a given distractor is determined. A distractor is selected by the least number of students eliminated and replaced by a NOTA-option. One study replaced the most frequently chosen distractor with the NOTA-option (Tollefson, 1987), which means that a distractor selected by the most number of students is eliminated and replaced by a NOTA-option. One study added NOTA as a fifth option to construct NOTA-items (Kolstad & Kolstad, 1991). This method is different from the others because no replacement method is applied. Instead, the NOTA-option is added as an extra response-option. Thus, this method makes the number of response-options increase. However, these studies did not give enough information about the advantage and disadvantage of the replacement method in terms of an empirical perspective.

Item statistics

Item difficulty.

Item difficulty is the proportion of students choosing the correct answer on any given item. The item difficulty index (p) is between 0 and 1. To illustrate, when 90% of test takers correctly answer a MCI, the value of p is 0.90. It means that most test takers correctly responded to the item. Therefore, it is an easy item. On the other hand, the item difficult index is small when a few number of test takers correctly respond to the item, such as 10% or 20%, which

presents that p is equal to 0.10 or 0.20, respectively. As a result, difficult items are indicated by the finding for a small item difficulty index, whereas easy items are represented by a big value.

Six out of seven empirical studies have provided item difficulty. One of the studies was a meta-analysis study by Knowles and Welch (1992), which reviewed approximately 20 previous studies about NOTA-items. A mean effect size (d) with a confidence interval was calculated ($d_{\text{NOTA-items}} = -0.17$, confidence interval = $(-0.21, 0.55)$) in order to conclude how NOTA-items or conventional-MCIs affected item difficulty. The findings showed no statistical difference in item difficulty between the NOTA-items and conventional-MCIs because the d -mean effect size is smaller than 0.2. In other word, using NOTA-items or conventional-MCIs did not appear to change the item difficulty index. However, this study's findings do not clearly generalize the results for the entire population. Therefore, more research is required.

The findings of other studies were different from the ones in the meta-analysis study by Knowles and Welch (1992) in terms of item difficulty, which indicated that the NOTA-items were statistically more difficult than conventional-MCIs (Tollefson, 1987; Rich & Johanson, 1990; Frary, 1991; Kolstad & Kolstad, 1991; Crehan et al., 1993). Namely, these studies indicated that NOTA-items made the items more difficult for test takers than did the conventional-MCIs.

The NOTA-option as a key or as a distractor was also empirically tested to identify how either role affected item difficulty. Three of the studies above compared the NOTA-option as key, NOTA-option as a distractor, and conventional-MCIs with regard to the item difficulty index (Tollefson, 1987; Rich & Johanson, 1990; Frary, 1991). The same results were found in these three studies. A NOTA-item with NOTA-option as key was the most difficult for test takers while conventional-MCIs were the easiest ones. To illustrate these studies more

technically, Tollefson (1987) compared means of NOTA as key, NOTA as distractor and conventional-MCIs ($\mu_{\text{nota-items-KEY}}=7.00$, $\mu_{\text{nota-items-FOIL}}=8.46$, $\mu_{\text{conventional-MCIs}}=9.35$, $T [12] =12$, $p<.05$). Frary (1991) investigated the mean of conventional-MCIs, NOTA as key, and NOTA as distractor ($\mu_{\text{conventional-MCIs}}=.66$, $\mu_{\text{nota}}=.61$; $\mu_{\text{nota-items-KEY}}=.58$; $\mu_{\text{nota-items-FOIL}}=.61$). Although Rich and Johanson (1990) used two methods, CTT and IRT, the findings were parallel to others: CTT ($\mu_{\text{nota-items}}=69.14$, $\mu_{\text{conventional-MCIs}}=71.22$, $t=2.14$, $p<.03$) and IRT ($\mu_{\text{nota-items}}=-18$, $\mu_{\text{conventional-MCIs}}=-70$, $t=2.81$, $p=.012$). Moreover, Rich and Johanson showed that item difficulty index for NOTA-items with NOTA as key and NOTA as distractors were similar to other studies (indices $\text{nota-items-KEY}=.68$; indices $\text{nota-items-FOIL}=.73$). In the light of these examples from multiple studies, the findings shows that the NOTA-items with NOTA as key have a small item difficulty index, which means that this type of item is the most challenging one for test takers.

To sum up, five studies presented that the NOTA-option makes items more difficult, while one study found that the NOTA-option did not significantly change item difficulty. In addition, three studies pointed out that using a NOTA-item with NOTA as a key was more difficult than both a NOTA-item with NOTA as distractor and conventional MCIs. More research is needed to accurately confirm that there is a significant impact of the NOTA-option on item difficulty. The current study will examine impact of NOTA-option on item difficulty regardless using NOTA option as key or distractor.

Item discrimination.

Item discrimination is assessed by using either the Pearson-product moment correlation, biserial, or point biserial correlation, or IRT statistics. Older measures, such as the Kelly D statistics are inappropriate but some are still used (Kingston & Kramer, 2013). One of the commonly used methods is the item-total correlation index, presenting how well the item

discriminates between students with high ability and low ability (Downing, 2005). The index is the degree of correlation between the scores on the item (0 or 1) and total test scores. The value of the correlation index usually ranges between -1 and 1 though some forms are limited by item difficulty. The item with a higher positive index is a better item. If the index of an item is negative, it is expected that the lower ability students would get more scores on the item than the higher ability students do. Therefore, an item with a negative item discrimination index was undesirable.

Five studies have empirically supported item discrimination over the past 25 years. Four found parallel results, showing no statistical difference between NOTA and conventional MCIs; however, one study found that NOTA items had more discrimination than conventional MCIs. Tollefson (1987) reports that item discrimination indices for NOTA as key, NOTA as distractor and conventional-MCIs were different ($\text{Median}_{\text{nota-items-KEY}}=.46$, $\text{Median}_{\text{nota-items-FOIL}}=.42$, $\text{Median}_{\text{conventional-MCIs}}=.60$). The findings were not significantly different. Frary (1991) indicated no significant difference between not only conventional-MCIs and the MCI with the NOTA-option ($\mu_{\text{conventional-MCIs}}=.32$, $\mu_{\text{nota}}=.32$) but also the NOTA-option as key and the NOTA-option as distractor ($\mu_{\text{nota-items-KEY}}=.32$; $\mu_{\text{nota-items-FOIL}}=.32$). Parallel results were found by the study by Crehan and Haladyna (1991) employed the same year, which indicated no evidence between NOTA-items and conventional-MCIs in the difference of item discrimination. Knowles and Welch (1992) reviewed 20 studies about NOTA-items by using meta-analysis. Although 11 studies calculated item-discrimination for NOTA-items in the past, only seven provided item statistics for item discrimination. After "d" effect size was calculated, the findings showed no difference between NOTA and conventional-MCIs in item discrimination ($d_{\text{NOTA-items}}= 0.01$, confidence interval = (-0.18, 0.20)). Therefore, four studies revealed that there was no statistical

difference between a conventional-MCI and NOTA-item; however, more studies are essential in order to use NOTA-items in the classroom and for large-scale assessment.

Rich and Johanson (1990) employed a method by combining methods from Lord (1953) and Henrysson (1971), in order to compare item discrimination of conventional MCIs and NOTA-items. They concluded that when the item difficulty index is at the moderate level (optimal level), item discrimination index is maximal. It is suggested that a NOTA-option can be used when the item difficulty index is higher than the optimal difficulty level. In other words, the NOTA-option can be used for easy items; however, these results are not generalized based on the result of only a single study. Therefore, more research is necessary. The current study will examine impact of NOTA-option on item discrimination regardless using NOTA option as key or distractor.

Test statistics.

Test reliability.

Test reliability is provided empirically in three out of seven studies. The findings from two of these found the same results while the other found conflicting results. Rich and Johanson (1990) calculated KR-20 reliabilities for the test with the NOTA-items and the test with the conventional-MCIs after the test was administered to 300 college students. The findings showed that reliability between the NOTA-items and conventional-MCIs were not significantly different ($KR-20_{NOTA}=0.835$, $KR-20_{conventional-MCIs}=0.797$). Kolstad and Kolstad (1991) found similar results to Rich and Johanson (1990) ($KR-20_{NOTA}=0.769$, $KR-20_{conventional-MCIs}=0.756$) by administering the test to 84 college students. Unlike other empirical studies about the NOTA-items, Tollefson (1987) examined the reliabilities for tests with the NOTA-items as key and the NOTA-items as a distractor and the conventional-MCIs after the tests were administered to 81

college students. The findings conflicted with the other two studies. The form with conventional-MCIs was more reliable than the test with the NOTA-items ($KR-20_{NOTA (Key)}=0.56$, $KR-20_{NOTA (Distractor)}=0.51$, $KR-20_{conventional-MCIs}=0.74$). As a result, over the past 25 years whereas the results of two studies revealed that there was no difference between NOTA-items and conventional-MCIs in test reliability, one found that conventional-MCIs increases test reliability. Therefore, further studies are required to identify whether the impact of NOTA-items on the test reliability is significant.

The number of options.

Although most empirical studies conducted during this same time frame related to this item guideline have been employed according to Haladyna et al. (2002), the findings of the studies were contradictory about the impact of the number of options on item characteristics (item difficulty and item discrimination) and test characteristics (test reliability and test validity) for 25 years. Further studies need to be done to determine if the number of options make a statistical difference to the four characteristics.

Elimination method.

Elimination methods have been reported in seven out of nine empirical studies, regarding the number of answer options of MCIs over the past 25 years. There are different types of elimination methods which are applied in these studies in order to build fewer options. The least-frequency-method was commonly applied, which means that the option with the least selected was deleted (Landrum et al., 1993; Delgado & Prieto, 1998; Abad, Olea & Ponsoda, 2001;

Shizuka et al., 2006). Two studies calculated the point-biserial correlation coefficient for every single option of an item and eliminated the option with the least discrimination (Owen & Froman, 1987; Trevisan et al., 1991). Additionally, a recent study randomly deleted an option of an item to construct the item with fewer options (Baghei & Amrahi, 2011).

Item statistics.

Item difficulty.

Six studies empirically have provided item difficulty for MCIs with four options and three options over the past 25 years. There are conflicting results found between MCIs with four options and three options among the studies. The item difficulty between MCIs with four options and three options were not statistically different in four studies although they applied a mix of elimination methods: the least frequency elimination method in three of the studies and the random deletion method in one study. Delgado and Prieto (1998) compared MCIs with three options and four options for three different forms (Form 1: $\mu_{\text{four-option}}=.65$, $\mu_{\text{three-option}}=.73$, $t=-1.58$, $p > 0.05$; Form 2: $\mu_{\text{four-option}}=.59$, $\mu_{\text{three-option}}=.62$, $t=-.50$, $p > 0.05$; Form 3: $\mu_{\text{four-option}}=.60$, $\mu_{\text{three-option}}=.66$, $t=-1.20$; $p > 0.05$). Abad and his colleagues (2001) found item difficulty (p for CTT and b for IRT) of the MCI by applying CTT and IRT ($p_{\text{four-option}}=.57$, $p_{\text{three-option}}=.59$; $b_{\text{four-option}}=.40$, $b_{\text{three-option}}=.34$). Shizuka and his colleagues (2006) reported that although MCIs with four options were slightly easier than MCIs with three options, they were not statistically different from each other ($\mu_{\text{four-option}}=.02$, $SD_{\text{four-option}}=0.93$, $\mu_{\text{three-option}}=.20$, $SD_{\text{three-option}}=0.81$; $t=-1.97$, $p=.06$, $df=26$, two tailed). A recent study which was carried out by Baghei and Amrahi (2011) evaluated the forms containing 30 MCIs with four options or three options each by using Rasch model ($\mu_{\text{four-option}}=0.09$, $\mu_{\text{three-option}}=-0.20$) and showed that the results were not statistically

different when fewer numbers of options were used. Further studies should be employed to determine any significant impact of three or four MCI options.

Two studies concluded that MCIs with three options were statistically more difficult than MCI with four options, which is a counterintuitive situation. One of these was investigated by Landrum and his colleagues (1993), in which they compared these types of items. ($\mu_{\text{four-option}}=82.0$, $\mu_{\text{three-option}}=86.8$; $t(143)=-5.70$; $p<.0001$). These results were parallel to the meta-analysis Rodriguez (2005) conducted. He examined 48 empirical studies from 1925 to 1999 in order to uncover the impact of the number of options in MCQ on psychometric characteristics. Twenty seven out of 48 studies related to achievement and attitude tests included an available report. The results presented that item difficulty slightly increased when the number of options reduced from four to three ($\mu_{\text{difference between four-option and three-option}}=0.04$, $p<0.05$).

As a result, four studies revealed that item difficulty was not significantly changed when the number of options of MCIs in a form decreased. Two other studies indicated that decreasing the number of options of a MCI makes that item more difficult. However, more research is essential to make conclusions about any impact of the number of options of a MCI on item difficulty.

Item discrimination.

Six empirical studies have investigated item discrimination for MCIs with four and three options over the past 25 years. Similar to the findings of item difficulty, mixed results were found for item discrimination. Item discrimination between MCIs with four and three options was not statistically different with three studies. Delgado and Prieto (1998) generated the three forms with MCIs with four and three options and compared them (Form 1: $r_{\text{four-option}}=0.36$, $r_{\text{three-option}}=0.34$, $t=0.51$, $p > 0.05$; Form 2: $r_{\text{four-option}}=0.28$, $r_{\text{three-option}}=0.28$, $t=-0.06$, $p > 0.05$; Form 3:

$r_{\text{four-option}}=0.33$, $r_{\text{three-option}}=0.29$, $t=-1.18$; $p>0.05$). Another study found parallel results although CTT and IRT methods are applied to investigate item discrimination (r for CTT and a for IRT: $r_{\text{four-option}}=0.37$, $r_{\text{three-option}}=0.37$; $a_{\text{four-option}}=1.03$, $a_{\text{three-option}}=1.01$) (Abad, Olea & Ponsoda, 2001). Shizuka and colleagues (2006) also found similar results for item discrimination for MCIs with four options and three-options ($r_{\text{four-option}}=0.31$; $r_{\text{three-option}}=0.29$).

Crehan and his colleagues (1993) carried out another study by adding a variable of the number of options to the one they administered before. Point-biserial (Pt-bis) discrimination indexes were calculated for 3-option regular, 3-option NOTA, 4-option regular and 4-option NOTA items. The findings showed that there was no different discrimination among these types of items ($\text{Pt-bis}_{\text{regular}(4)}=.36$, $\text{Pt-bis}_{\text{NOTA}(4)}=.37$, $\text{Pt-bis}_{\text{regular}(3)}=.33$, $\text{Pt-bis}_{\text{NOTA}(3)}=.33$).

Two studies provided statistically significant evidence for the MCIs, deducing that item discrimination for MCI with three-options is higher than the MCIs with four-options. Baghei and Amrahi (2011) examined the discrimination power of options in three steps: their average measures, outfit mean squares, and point-measure correlation. They concluded that the MCIs with three options had a higher discrimination power. Rodriguez (2005) conducted meta-analyses and found that item discrimination of MCIs with three-options slightly higher than the one with four-options ($r_{\text{difference between four-option and three-option}}=.03$, $p<0.05$).

Consequently, while four studies indicated that item discrimination was not significantly changed when the number of options of MCIs in a form decreased, two studies presented that item discrimination increased significantly. Therefore, more empirical studies are necessary to confirm any the impact of the number of option of a MCI on item discrimination.

Test characteristics.

Test reliability.

Four studies have been conducted to investigate the impact of the number of options on test reliability over the past 25 years. Mixed results were found for test reliability between forms containing MCIs with four options and those containing three options.

Two studies revealed that the number of options did not have statistically significant impact on test reliability. Delgado and Prieto (1998) examined three forms which consisted of the MCIs with four and three options, and they calculated Spearman-Brown adjusted reliability coefficients (r) for each set of item groups in each form (Form 1: $r_{\text{four-option}}=0.78$, $r_{\text{three-option}}=0.70$; Form 2: $r_{\text{four-option}}=0.64$, $r_{\text{three-option}}=0.63$; Form 3: $r_{\text{four-option}}=0.68$, $r_{\text{three-option}}=0.76$). Another study utilized Cronbach's Alpha reliability coefficients (r) for the forms including MCIs with four or three options (Form 1: $r_{\text{four-option}}=0.79$; Form 1: $r_{\text{three-option}}=0.76$) (Baghei & Amrahi, 2011).

One study found that test reliability changed for the forms consisting of MCIs with fewer options. Rodriguez (2005) investigated test reliability by applying meta-analyses for the past studies and found that test reliability slightly increased when the forms with three-options were employed ($r_{\text{difference between four-option and three-option}}=.02$, $p<0.05$).

Only one study investigated test reliability for low, average, and high ability students (Trevisan et al., 1991). Reliability coefficients decreased when the number of options decreased from four to three for low ability students ($\chi^2(2, N=97)=9.21$, $p \leq 0.05$). However, the reliability coefficient did not find a statistical difference for average ability students ($\chi^2(2, N=89)=2.97$, $p > 0.05$) and high ability students ($\chi^2(2, N=96)=0.29$, $p > 0.05$).

To sum up, two studies found parallel results, showing that test reliability was not significantly affected when a form consisted of MCIs with fewer number options. One study indicated that test reliability decreased whereas another study found mixed results for the students with different abilities when the forms with fewer options were used. Therefore, more research is needed to conclude a significant impact of MCIs with fewer options on test reliability.

Test validity.

There is only one study which investigated the impact of the number of options on test validity over the past 25 years. The findings of the study were akin to the impact of the number of options on test reliability (Trevisan et al., 1991). Validity coefficient as criterion validity was calculated in this study by asking students to state their GPA. The correlation between students' GPA and the test administered in this study showed the validity-coefficient. The number of options negatively affected validity-coefficient ($\chi^2 (2, N=97) = 54.19, p \leq 0.05$) for low ability students. However, there was no statistical difference in the validity coefficient for average ability students ($\chi^2 (2, N=89) = 3.95, p > 0.05$) and high ability students ($\chi^2 (2, N=96) = 2.07, p > 0.05$). In conclusion, a 3-option test form is most valid for the student with or without considering their ability, which did not support the findings of previous studies. However, further study is necessary to confirm that this guideline does not affect test validity.

Option Replacement and Elimination with Item Response function

To decrease the number of options for a MCI, Item Response Function (IRF) is employed. IRF is the mathematical model for relationship between abilities of examinees and their probability for correct response of an item. This model is basically based on the assumption of unidimensionality, defined as one construct of the test of interest measures (Hambleton,

Swaminathan, & Rogers, 1991, p.7); monotonicity of IRF, the relationship between abilities of examinees and their probability for correct response always increase monotonically; and parameter invariance, “the parameters that characterize an item do not depend on the ability distribution of the examinees and the parameter that characterizes an examinee does not depend on the set of test items” (Hambleton, Swaminathan, & Rogers, 1991, p.18).

In common item response models, IRFs are assumed to be monotonically increasing, having an “S” shape. However, each distractor’s characteristics might not be monotonically increasing. Instead of parametric IRF, non-parametric IRF is employed to investigate the characteristics of distractors.

The intended data does not always fit the model in terms of monotonicity of IRF (Ramsay, 1991). In other words, the relationship between abilities of examinees and their probability for correct response does not always increase monotonically. In this circumstance, nonparametric IRF is applied rather than parametric IRF.

Nonparametric smooth regression method.

The most widely used nonparametric IRF model is the Gauss kernel smoothing regression method in order to generate a graph of relationship between an examinee’s ability and the examinee’s probability to get a correct answer (Lee, 2007). The Gauss kernel smoothing regression method is based on the principle of local averaging (Altman, 1992; Eubank, 1988; Simonoff, 1996). Specifically, uniform local averaging is employed in this method. Assume that (x, y) is a point of regression function. While x is an independent variable, y is a dependent variable which coincides with the variable of x . Based on the principle of uniform local averaging, y is the average of y_i s which is defined as the average of x_i s which is less than h unit

from x . Moreover, each value of x_i in the same local has different weights in terms of their distance to x . The following equation shows the weight of x_i :

$$w_j = \frac{K((x-x_i)/h)}{\sum_{j=1}^n K((x-x_j)/h)} \quad (1)$$

Where $i=1, 2, \dots, n$ examinees, $j=1, 2, \dots, n$ items, and h is bandwidth, which is $h = 1.1N^{-\frac{1}{5}}$ in the TestGraf, which is applied in this study. K is the Gaussian kernel smoothing function indicated below:

$$K(u) = e^{-\frac{u^2}{2}}. \quad (2)$$

When this function is applied for all individuals (x_i, y_i) , the following equation shows the complete definition of the Gauss kernel smoothing estimate function:

$$g(x) = \sum_{i=1}^n w_i y_i \text{ where } w_i = \frac{K((x-x_i)/h)}{\sum_{j=1}^n K((x-x_j)/h)}. \quad (3)$$

The value of y_i is equal to 1 if the item is answered correctly, and 0 if not. This formula is rewritten in terms of options of a MCI. It means that it shows the relationship between an examinee's ability and his response to get the correct answer to an item. While the independent variable is an examinee's ability (θ), the dependent variable is the probability of choosing option m for item j , $p_{jm}(\theta)$. The following equation shows the relationship:

$$p_{jm}(\theta) = \sum_{i=1}^n w_i y_{jmi} \text{ where } w_i = \frac{K((\theta-\theta_i)/h)}{\sum_{j=1}^n K((\theta-\theta_j)/h)}. \quad (4)$$

The value of y_{jmi} is equal to 1 if the examinee I choose the option m of item j , and 0 if not. In this equation, θ_i is unobservable variable, which is known as latent variable. Therefore, it is replaced with $\hat{\theta}$:

$$Pp_{jm}(\theta) = \sum_{i=1}^n w_i y_{jmi} \text{ where } w_i = \frac{K((\theta-\hat{\theta}_i)/h)}{\sum_{j=1}^n K((\theta-\hat{\theta}_j)/h)}. \quad (5)$$

The Number of Item Solution Steps with bi-factor IRT Model

The bi-factor IRT model is appropriate model for the data because each item loads the general dimension factor and only one sub-dimension factor in this model (Gibbons & Hedeker, 1992). Also in this model, the data fits the multidimensional IRT model even though it fits the unidimensional IRT model (Reise, Morizot, & Hays, 2007). Therefore, the bi-factor IRT model is an alternative model of the unidimensional IRT model (Gibbons & Hedeker, 1992). By using the bi-factor IRT model, the amount of item variance is due to general factors and sub-group factors (Reise, Morizot, & Hays, 2007). The bi-factor IRT model addresses the question that the amount of information for each item comes from general and sub-group factors.

Throughout this process, the loading matrix which shows the degree of each item's loading on the general factor and two sub-group factor was generated. All items are loading to general factor and one of sub-group factors, which is an assumption of the bi-factor IRT model. Figure 1 illustrates how six items works with bi-factor IRT model. In this model, the first three items are loaded to sub-group factor S1 while others are loaded to sub-group factor S2. All of six items also are loaded to general factor G.

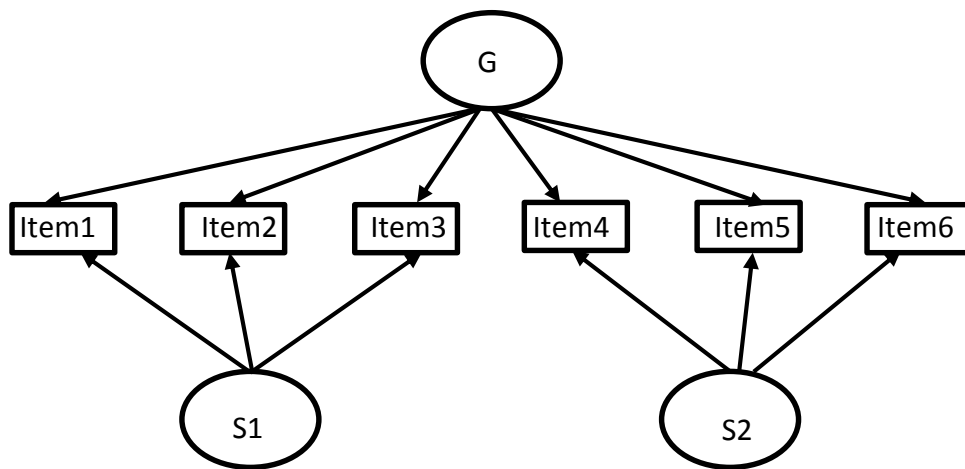


Figure 1. An example of bi-factor model.

Summary

This chapter examined the item writing guidelines which were cited by Haladyna et al. (2002). Additionally, two of the guidelines and their impact on item characteristics and test characteristics were provided according to the literature. One concerns the number of options of MCIs in a form while the other one addressed “None of the above” as an option. We also illustrated how to analyze MCIs’ options and determine the number of solution steps of the MCI.

CHAPTER 3

Methodology

This section consists of four subsections: participants, instrument development, final instrument and data analysis.

Participants

Seventh and eighth grade students from Turkey and the U.S. participated in two test administrations, the pilot and the final. The convenience sampling method was applied for two administrations. For the pilot administration in the spring semester of 2012, 1,130 7th grade students participated from sixteen schools in Turkey, while 100 students were from only one school in the U.S. The final test was administered to 1,082 7th and 8th grade students in Turkey in the fall semester of 2012 and 585 7th and 8th grade students in the U.S. during the spring semester of 2013. In both administrations, only students' responses to math items were collected without identifying their prior ability and demographic information.

Instrument Development

The purpose of this section is to document the development of the instruments used in this research. One of the important aspects of test development is to determine the construct or constructs of the test and appropriate content for the construct. Since the construct of the test is defined as what the test measures, it is an abstract term. When appropriate content of the construct is applied in the test, the abstract construct becomes the concrete term by providing the significant test results (Kingston & Kramer, 2013). One of the ways to provide test content in larger detail is to use a test blueprint, which contains a great number of elements, such as content area, number of items in each content area, total number of items in the test, educational level of

examinees, and duration of test administration. Therefore, after the content area of the test designed in this study was chosen based on the Common Core State Standard (CCSS), all items were written based on this content area. As a result, test validity increased in terms of content area since content-valid items are applied in the test, which enables the systematic errors to be controlled.

Fifty eight items were developed based on one domain of the CCSS mathematics, “Equation & Expressions”, the reliable items among the valid items are chosen. Two forms (Form A and Form B) with 29 items each were constructed so participating students could answer all items in one class period. Two forms contained parallel items, which have the same content and rationale of distractors. In other words, two parallel items were constructed, and one of them was placed in Form A while the other one was placed in Form B.

Students’ responses for each individual item were used to determine the items’ characteristics, item difficulty and item discrimination. Thirty out of 58 items with higher discrimination and middle item difficulty were selected in order to control the random errors in the final instrument. Consequently, 30 valid and reliable MCIs with four options were designed for use in the final instrument. In the next stage, we generated 3 forms based on the 30 MCIs with four options: 10 with four options, 10 with three options (once an option was removed), and 10 with NOTA options (once an option was replaced with NOTA).

While one of four options for a MCI was eliminated to construct a MCI with three options, it was replaced by a NOTA option to construct a MCI with a NOTA option. One of the biggest issues was determining which option to select for elimination or replacement. Although different elimination and replacement methods are used in the previous studies, the Gauss Kernel smoothing regression method was applied in this study. In this method, the option of a MCI to be

eliminated or replaced was based on the Item characteristic Curve (ICC), the relationship between abilities of examinees and their probability for correct response always increases monotonically (Hambleton, Swaminathan, & Rogers, 1991). ICC provides a good psychometric quality when increasing monotonically. However, ICC does not increase monotonically for some items all the time. In other words, ICC of an item decreases at some points where one of the options of the same item increases. Therefore, this option is eliminated to allow ICC to increase monotonically. Therefore, this method is applied to construct 10 MCIs with three options and 10 MCIs with NOTA options for use in the final administration.

Procedure of item writing and test administration.

Fifty-eight multiple-choice math items with four options were written using on the guidelines from Haladyna (2002). One of the options for each item was the key whereas others were the distractors. These questions were written based on one of the domains, a large group of related standards, in the Common Core State Standard (CCSS), “expressions and equations”, a foundational and critical unit for the students because the math knowledge taught in the next grades is based on this unit. In other words, students cannot understand the math knowledge in the next grade without understanding “expressions and equations.” For example, they cannot solve the multi-step word problems and advanced-level math problems without using “expression and equations”. Seventh grade is the first time students start to learn “expression and equation” by using one variable. Therefore, 7th grade is the first step to learn high-level math and how to solve high-level math problems.

Each domain consists of four standards. Different numbers of multiple-choice items were written for each standard. Table 2 shows the relationship between the number of items and the standard.

Table 2. *The relationship between the common core standards and the number of items*

The Common Core State Standard	The Number of Items
7.EE.1. Apply properties of operations as strategies to add, subtract, factor, and expand linear expressions with rational coefficients.	21
7.EE.2. Understand that rewriting an expression in different forms in a problem context can shed light on the problem and how the quantities in it are related. <i>For example, $a + 0.05a = 1.05a$ means that “increase by 5%” is the same as “multiply by 1.05.”</i>	10
7.EE.3. Solve multi-step real-life and mathematical problems posed with positive and negative rational numbers in any form (whole numbers, fractions, and decimals), using tools strategically. Apply properties of operations to calculate with numbers in any form; convert between forms as appropriate; and assess the reasonableness of answers using mental computation and estimation strategies. <i>For example: If a woman making \$25 an hour gets a 10% raise, she will make an additional 1/10 of her salary an hour, or \$2.50, for a new salary of \$27.50. If you want to place a towel bar 9 3/4 inches long in the center of a door that is 27 1/2 inches wide, you will need to place the bar about 9 inches from each edge; this estimate can be used as a check on the exact computation.</i>	15
7.EE.4. Use variables to represent quantities in a real-world or mathematical problem, and construct simple equations and inequalities to solve problems by reasoning about the quantities.	
a. Solve word problems leading to equations of the form $px + q = r$ and $p(x + q) = r$, where p , q , and r are specific rational numbers. Solve equations of these forms fluently. Compare an algebraic solution to an arithmetic solution, identifying the sequence of the operations used in each approach. <i>For example, the perimeter of a rectangle is 54 cm. Its length is 6 cm. What is its width?</i>	12
b. Solve word problems leading to inequalities of the form $px + q > r$ or $px + q < r$, where p , q , and r are specific rational numbers. Graph the solution set of the inequality and interpret it in the context of the problem. <i>For example: As a salesperson, you are paid \$50 per week plus \$3 per sale. This week you want your pay to be at least \$100. Write an inequality for the number of sales you need to make, and describe the solutions.</i>	
	58 (total)

Fifty eight MCIs were written, as shown in Table1, to build two forms (Form-A and Form-B) for the pilot study. Each is composed of 29 multiple-choice math items. Therefore, students can easily solve one form per one-classroom period, approximately 40-50 minutes long. In Turkey, Form A was taken by 656 students from seven junior high schools in one-classroom period while Form B was taken by 474 students from seven junior high schools in one-classroom period. In the U.S., 100 students from only one junior high school took both Form A and Form B in a two-classroom period.

After the pilot data was collected, the responses of each individual student were manually entered from paper-pencil test to excel-spreadsheet. Item discrimination (r) and item-difficulty (p) for each individual item were calculated for every form in Turkey and the U.S. The most common way to calculate item discrimination is the item-total correlation index, presenting how well the item discriminates between students with high ability and low ability (Downing, 2005). Item difficulty is the proportion of students choosing the correct answer. Table 3 indicates the values of item-total correlation and item difficulty for each item among the different forms.

Before analyzing data, the meaning of superscript of the index is explained. For example, suppose a superscript like (a, b) . a , the left number, shows to which standard the item belongs while b , the right one, to which group the item belongs. Each standard has a different number of groups. Each group has parallel items, which have the same content and rationale of distractors, and require the students to take the same steps to calculate the item. However, the numbers which are used in the stem and options are different. To illustrate, twenty-one MCIs were written in four groups $((1,1), (1,2), (1,3), (1,4))$ in standard-1, *7.EE.1*. In other words, four groups are formed in standard-1.

Table 3. *The values of item discrimination and item difficulty for each item in pilot data*

Questions	Form A (TUR)		Form A (US)		Selected	Form B (TUR)		Form B (US)		Selected
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>		<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	
Q1	0.60 ^{1,1}	0.55	0.38 ^{1,1}	0.76	yes	0.43 ^{1,1}	0.59	0.63 ^{1,1}	0.66	yes
Q2	0.64 ^{1,1}	0.53	0.35 ^{1,1}	0.76	no	0.46 ^{1,1}	0.55	0.48 ^{1,1}	0.71	no
Q3	0.61 ^{1,1}	0.54	0.44 ^{1,1}	0.74	yes	0.41 ^{1,2}	0.64	0.68 ^{1,2}	0.62	no
Q4	0.42 ^{1,2}	0.66	0.56 ^{1,2}	0.69	no	0.43 ^{1,2}	0.65	0.69 ^{1,2}	0.71	no
Q5	0.40 ^{1,2}	0.71	0.58 ^{1,2}	0.72	no	0.46 ^{1,2}	0.64	0.77 ^{1,2}	0.67	no
Q6	0.38 ^{1,2}	0.74	0.56 ^{1,2}	0.73	no	0.52 ^{1,3}	0.54	0.61 ^{1,3}	0.74	yes
Q7	0.56 ^{1,3}	0.50	0.55 ^{1,3}	0.82	no	0.52 ^{1,3}	0.55	0.63 ^{1,3}	0.84	no
Q8	0.52 ^{1,3}	0.52	0.59 ^{1,3}	0.81	yes	0.52 ^{1,3}	0.54	0.60 ^{1,3}	0.87	no
Q9	0.61 ^{1,4}	0.58	0.47 ^{1,4}	0.67	yes	0.51 ^{1,4}	0.54	0.68 ^{1,4}	0.64	no
Q10	0.56 ^{1,4}	0.63	0.60 ^{1,4}	0.67	yes	0.49 ^{1,4}	0.59	0.73 ^{1,4}	0.65	yes
Q11	0.55 ^{2,1}	0.35	0.70 ^{2,1}	0.57	no	0.53 ^{1,4}	0.59	0.64 ^{1,4}	0.64	no
Q12	0.55 ^{2,1}	0.36	0.68 ^{2,1}	0.59	yes	0.42 ^{2,1}	0.50	0.60 ^{2,1}	0.65	yes
Q13	0.53 ^{2,1}	0.47	0.48 ^{2,1}	0.71	yes	0.38 ^{2,1}	0.33	0.52 ^{2,1}	0.65	no
Q14	0.50 ^{2,2}	0.45	0.53 ^{2,2}	0.49	yes	0.42 ^{2,2}	0.44	0.33 ^{2,2}	0.49	no
Q15	0.43 ^{2,2}	0.43	0.51 ^{2,2}	0.63	no	0.42 ^{2,2}	0.45	0.40 ^{2,2}	0.56	yes
Q16	0.53 ^{3,1}	0.59	0.48 ^{3,1}	0.86	yes	0.47 ^{2,2}	0.37	0.39 ^{2,2}	0.60	yes
Q17	0.63 ^{3,1}	0.6	0.42 ^{3,1}	0.65	yes	0.57 ^{3,1}	0.31	0.40 ^{3,1}	0.67	no
Q18	0.41 ^{3,1}	0.52	0.48 ^{3,1}	0.59	no	0.66 ^{3,1}	0.36	0.46 ^{3,1}	0.70	yes
Q19	0.56 ^{3,2}	0.50	0.49 ^{3,2}	0.58	yes	0.60 ^{3,1}	0.38	0.50 ^{3,1}	0.66	no
Q20	0.62 ^{3,2}	0.52	0.63 ^{3,2}	0.68	yes	0.50 ^{3,2}	0.39	0.42 ^{3,2}	0.53	yes
Q21	0.30 ^{3,3}	0.30	0.45 ^{3,3}	0.41	yes	0.56 ^{3,2}	0.34	0.29 ^{3,2}	0.45	no
Q22	0.40 ^{3,3}	0.39	0.51 ^{3,3}	0.37	yes	0.43 ^{3,3}	0.26	0.38 ^{3,3}	0.45	no
Q23	0.25 ^{3,3}	0.34	0.52 ^{3,3}	0.40	no	0.50 ^{3,3}	0.34	0.28 ^{3,3}	0.43	yes
Q24	0.47 ^{4,1}	0.64	0.48 ^{4,1}	0.84	yes	0.43 ^{4,1}	0.64	0.50 ^{4,1}	0.66	no
Q25	0.54 ^{4,1}	0.5	0.59 ^{4,1}	0.83	yes	0.50 ^{4,1}	0.47	0.50 ^{4,1}	0.49	yes
Q26	0.55 ^{4,1}	0.53	0.60 ^{4,1}	0.84	no	0.50 ^{4,1}	0.47	0.42 ^{4,1}	0.56	no
Q27	0.20 ^{4,2}	0.19	0.37 ^{4,2}	0.45	no	0.30 ^{4,2}	0.42	0.42 ^{4,2}	0.45	yes
Q28	0.34 ^{4,2}	0.33	0.43 ^{4,2}	0.52	yes	0.25 ^{4,2}	0.30	0.42 ^{4,2}	0.44	no
Q29	0.24 ^{4,2}	0.26	0.38 ^{4,2}	0.51	no	0.35 ^{4,2}	0.28	0.35 ^{4,2}	0.49	yes

(x,y): x for the standard; y for the group

At least four MCIs were generated for each group, respectively, five MCIs for (1,1), six MCIs for (1,2), five MCIs for (1,3), and five MCIs for (1,4).

If the number of MCIs in each group is even (four or six), the first half is in Form-A while the second half of them is in Form-B. For example, there are six MCIs for (1,2), three of them in Form-A and the other three in Form-B. If the number of items in each group is odd (five), while three of them are in Form-A, two of them are in Form-B, or vice versa. To illustrate, five MCIs for (1,1) were written, three of them in Form-A and two of them in Form-B.

Item selection.

To generate the final data, thirty MCIs were selected from the pilot data by using item discrimination and item difficulty index for each item from Table 3. In terms of research design, three MCIs were selected from each group because one of them was kept without changing while the others will be rewritten in two versions: four option with NOTA-option and, three option, respectively. In addition, the MCIs with NOTA-options were key-balanced. An example of each MCI version used in this study is given in Table 4 by using the MCIs from group-2 in standard-1, (1,2).

To eliminate the items from the same group in order to generate three items for the final test, the item-total correlation index (r) was a major consideration. The item with lower r in the same group was eliminated. If r was similar among the items in the same group, then the item with high item difficulty index (p) was eliminated from the test. For example, in Table 3, Q14 (2, 2) in Form B in Turkey and the U.S. has the lowest r , 0.33, among the items in the same group. Thus, it was eliminated. Q13 (2, 1) had similar r as the other items in the same group, however it

was eliminated due to its higher p value because an item with higher p decreases test variance. Thus, three items were selected in the same group in order to generate the items in different versions. Table 3 shows the selected and deleted items from Form A and Form B. 30 selected items were used in the final administration.

Table 4. *The example of the items in different versions in the final test*

Final Test	Version
Which expression could be used to find 9 less than 12 times x ? A. $9 - 12x$ B. $12x - 9$ C. $x(12 - 9)$ D. $12(x - 9)$	Four option (the base item)
Which expression could be used to find 4 less than 5 times x ? A. $4 - 5x$ B. $5x - 4$ C. $x(5 - 4)$ D. None of the Above	Four option with NOTA-options
Which expression could be used to find 8 less than 10 times x ? A. $8 - 10x$ B. $10x - 8$ C. $x(10 - 8)$	Three option

Option elimination with the Gauss Kernel smoothing regression method.

To eliminate an option of a MCI, the graph of each option of the MCI is made to show the relationship between examinees' probability of that response and their ability. Data from the same forms were combined across the two countries. In other words, while Form-As from

Turkey (n=656) and the U.S.A. (n=100) are combined, Form-Bs from Turkey (n=474) and the U.S.A. (n=100) were combined. Therefore, the number of students increased in Form-A and Form-B: 756 students and 574 students, respectively. Twenty-nine graphs for each form were generated because each form involves 29 MCIs. To generate the graphs, TestGraf was employed. However, TestGraf does not allow the graphs to be edited. Thus, the output file which consists of the variables for the Kernel Smoothing Method (KSM) was obtained. This file also involves some variables and texts which are not related to KSM. Therefore, to get only variables which correspond to the KSM, one Fortran Program was written to generate a new output file for STATA. Then, STATA was used to build the graphs. Graphs for all items are shown at Appendix B. One of the graphs is shown in Figure 2.

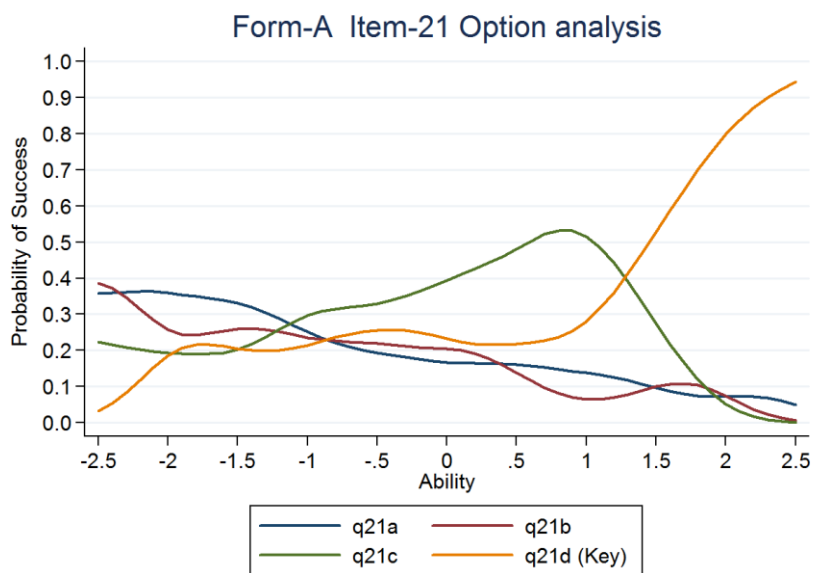


Figure 2. Option Analysis of Item 21 in Form-A using Gauss Kernel Smoothing Regression Method.

The key of Item 21 is option D, which is shown with an orange color line. However, the probabilities that low ability students would select this option are lower than the other options. In addition, the probabilities that middle ability students would select the key are lower than option C. For the high ability students, the probability of selecting the key is higher than the other options. In terms of item reliability, the options which are the distractors make item reliability decrease. To illustrate, it is expected that the graph of key, or ICC, increases monotonically. However, the graph of key for Item-21 both increases and decreases. In addition, some location decreases in the one makes the other options increase, which decreases item validity. For instance, between 0 and 1, the graph of key decreases slightly, and option C increases. Therefore, option C is the best option to be eliminated from the item.

This elimination method does not work for some items because there is no location showing that the graph of their distractors increases when the graph of key decreases. Figure 3 shows an example of this situation. The key is shown by the green line. It is monotonically increasing for all locations. The graph of each distractor shows that they are both increasing and decreasing. However, when the graph of each distractor increases, the graph of keys does not decrease.

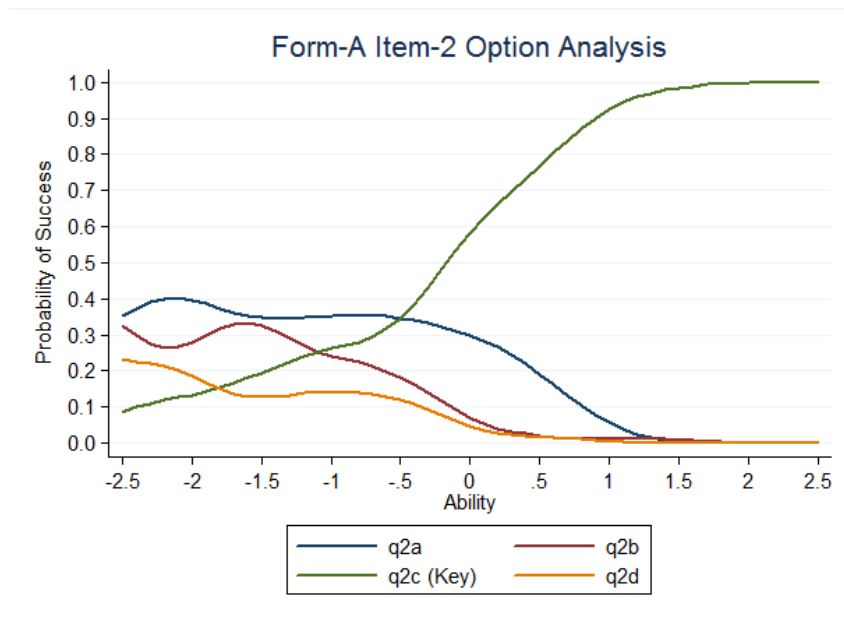


Figure 3. An example of the item in which content of options as elimination method is applied.

Thus, one of the options is eliminated in terms of content of options. In other words, the weak rationale of a distractor is deleted. Table 5 below shows what types of methods, graph or content, was applied in order for all individual items to eliminate one of options.

Table 5. Types of method to eliminate the options

Form A		Form B	
# of Item	Elimination Method	# of Item	Elimination Method
3	Content	1	Content
10	Graph	6	Graph (key)
13	Content	8	Content
16	Content	10	Content
20	Graph	12	Graph (key)
21	Graph	15	Graph
24	Graph	16	Content
25	Content	18	Graph
		20	Graph
		23	Graph
		27	Graph (key)
		29	Content

Final Instrument

30 multiple choice math items (MCMI) were selected to use in the final administration. MCMI are written in 3 different formats: MCMI with four options, three options, and with the NOTA option. Therefore, there are three forms consisting of the same number of MCMI with different formats. To illustrate, 10 MCMI with four options are located in Form A, 10 MCMI with three options are located in Form B, and 10 MCMI with NOTA options are located in Form C. After the forms are located in the order of Form A, Form B, and Form C, respectively, they are administered to students. Therefore, MCMI with four options are numbered as #1—10, MCMI with three options numbered as #11—20, and MCMI with NOTA options are numbered as #21—30. Items in different forms are parallel to each other in terms of their content. However, parallel items are randomly located in each form. For instance, three parallel items are located at #1, #16, and #23, each corresponding to a different form.

Data Collection.

The forms in the pilot study were only administered to 7th grade students since the content area of the forms, “Equation and Expressions,” is a 7th grade standard in CCSS. However, 7th grade math teachers in most of the schools had not completely covered the content of the form at the time the final administration was applied. Therefore, the paper-pencil form was administered to 7th and 8th grade students in the final administration at the beginning of spring semester in 2013 in Turkey as well as during spring semester in 2013 in the U.S.A. Cluster sampling is applied since the form is administered to all 7th and 8th grade classes. Convenience sampling method is used to choose the schools in Turkey and USA. One thousand eighty two students from five different schools in a city were chosen in Turkey, while 585 students from

five different schools in three different states were chosen in the U.S.A. Students individually responded to each item by selecting the appropriate answer option on the paper form in one class period, approximately 40-50 minutes in length. After the current teacher gathered each paper form from the students at the end of the class, they brought all the forms to the office of the school administrator. After gathering all forms, the responses of the items were manually entered on the computer for the data analyses.

Determining number of steps for items.

The items in this research were written based on the CCSS. Four standards (7.EE.1, 7.EE.2, 7.EE.3, and 7.EE.4) were applied while writing the questions as shown in Table 1. Two of them (7.EE.1 and 7.EE.2) require one-step problem of equation and expression. This means that each individual problem based on these standards had a one-step solution, therefore, evaluating each student's skill level. On the other hand, the problems based on other two standards (7.EE.3 and 7.EE.4) had more than one-step real life word problems (multi-step solutions). Table 6 shows the examples of MCIs with one-step and multi-step solutions.

Table 6. *Problems with one-step and multi-step solution*

One-step solution	Multi-step solutions
Which expression could be used to find 3 times 5 more than x ?	Jay has \$80. He uses 60% of his money to buy a pair of shoes. Then, he spends $\frac{1}{4}$ of the remaining money to buy a gift for his friend. How much money does he pay for the gift?
A. $3x + 5$	A. \$12
B. $5x + 3$	B. \$10
C. $3(x + 5)$	C. \$8
D. $(5 + 3)x$	D. \$5

Students may be able to find correct answer without taking any pen and paper at one-step problem, which is written based on 7.EE.1. However, they need to make several calculation steps

which are more likely to require pen and paper to find correct answer of multi-step problem, which is aligned to 7.EE.3. Moreover, students have extra effort to convert the word problem to mathematical expressions for multi-step problems.

Based on the content, the number of items with a one-step solution and multi-step solutions in final instrument are the same, 15 each. In terms of bi-factor IRT model, all 30 items in this study are related to mathematics ability as general factor. While 15 items with one-step solutions are loaded one of sub-group factors, other 15 items with multi-step solutions are loaded to other sub-group factor. When the model fits are acceptable, we mark each item as “one-step” or “multi-step”. After this process, we examine the research questions with particular methods, which are described below.

Data Analysis

For research question 1, item difficulty and item discrimination for each item for NOTA-items and conventional-MCIs with four options were calculated. Also, the number of solution steps of each item was determined. Then, a multiple regression model was conducted to evaluate the effects of item type (4 vs. NOTA) and the number of solution steps on item difficulty and item discrimination. Moreover, the interaction effect between item type and the number of solution steps was evaluated to find whether there was a difference in the mean of item difficulty and item discrimination among the two types of items when the number of solution steps varied.

For research question 2, reliability coefficients for each set of item group (four option MCIs with one-step solution, four option MCIs with multi-step solutions, NOTA-MCIs with one-step solution, and NOTA-MCIs with multi-step solutions) were analyzed by applying the

Kuder-Richardson Formula (KR-20), which was a particular form of Cronbach's α (Cortina, 1993). Furthermore, standard error of each reliability coefficient is calculated to examine whether the reliability of each item set was statistically different.

For research question 3, item difficulty and item discrimination for each item for MCIs with three options and conventional-MCIs with four options were calculated. Also, the number of solution steps of each item was determined. Then, a multiple regression model was conducted to evaluate the effects of item type (4 vs. 3) and the number of solution steps on item difficulty and item discrimination. Moreover, the interaction effect between item type and the number of solution steps was evaluated to find whether there was a difference in the mean of item difficulty and item discrimination among the two types of items when the number of solution steps varied.

For research question 4, reliability coefficients for each set of item group (four option MCIs with one-step solution, four option MCIs with multi-step solutions, three option MCIs with one-step solution, and three option MCIs with multi-step solutions) were analyzed by applying the Kuder-Richardson Formula (KR-20), which was a particular form of Cronbach's α (Cortina, 1993). Additionally, standard error of each reliability coefficient was calculated to examine whether the reliability of each item set was statistically different.

CHAPTER 4

Results

In this section, we determine the number of solution steps for each item using bi-factor IRT model and then investigate six research. Three software packages, IRTPRO 2.1 (Cai, Thissen & du Toit, 2011), STATA (StataCorp, 2013) and SPSS (Statistics, 2012), are used. IRTPRO was used to confirm the existence of separate factors based on number of steps needed to solve each item while STATA and SPSS were used to investigate four research questions.

Confirming the Existence of Separate Factors Based on Number of Steps with Bi-Factor IRT Model

This process includes several steps: a) generating unidimensional IRT model b) generating bi-factor IRT model c) Comparing unidimensional and bi-factor IRT models. Unidimensional IRT model, which not include any sub-group factor, has only a general factor where all items are loaded. Figure 4 shows an example of unidimensional IRT model with six items.

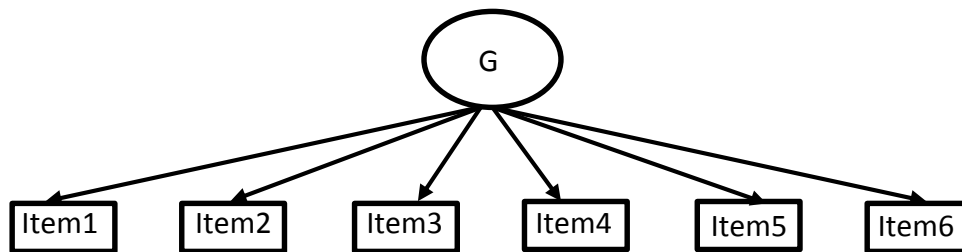


Figure 4. An example of unidimensional model.

In bi-factor IRT model, all items are loaded to general factor (G). In addition, the items with a one-step solution are loaded to one of the sub-factors of S1, the items with multi-step solutions are loaded to other sub-factor (S2), as shown Figure 5.

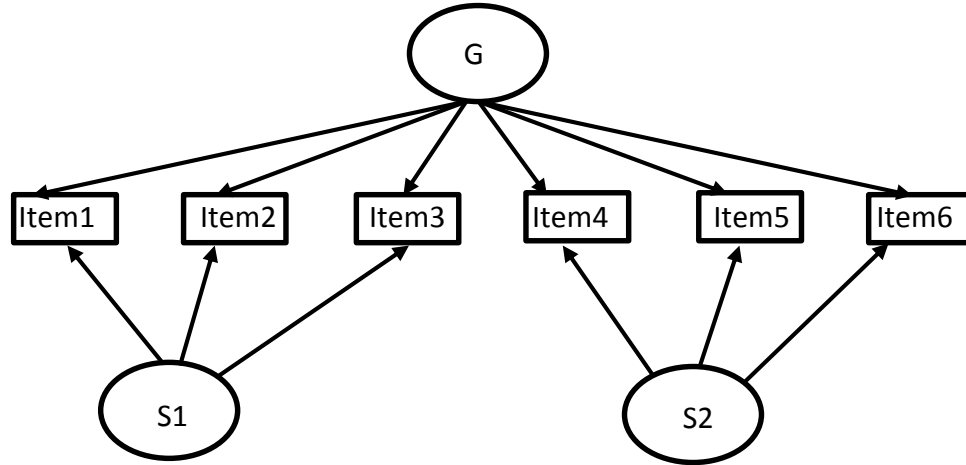


Figure 5. An example of bi-factor model.

In this study, we generates unidimensional IRT model and bi-factor IRT model for 30 items. Table 7 shows such models, their fit statistics (-2loglikelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC), and items with one-step solution and multi-step solutions.

Table 7. *Unidimensional and bi-factor model*

	Items loading to Factors			Model Fit		
	General Factor (G)	One-Step (S1)	Multi-steps (S2)	-2loglikelihood	AIC	BIC
Unidimensional Model	all items	-	-	59417.26	59537.26	59862.39
Bi-factor IRT Model	all items	1, 2, 4, 7, 10 12, 13, 15, 17, 19, 22, 24, 26, 28, 29	3, 5, 6, 8, 9, 11, 14, 16, 18, 20, 21, 23, 25, 27, 30	58680.92	58860.92	59348.61

Table 7 reveals that the bi-factor IRT model fits the data significantly better than unidimensional model. All items loading values to the factor G, S1 and S2 in bi-factor IRT model are shown at Appendix C.

Table 8. *Summary of Item Characteristics*

Items	Step Solutions		Option Format		
	One-Step	Multi-Step	Four	Three	NOTA
#1	x		x		
#2	x		x		
#3		x	x		
#4	x		x		
#5		x	x		
#6		x	x		
#7	x		x		
#8		x	x		
#9		x	x		
#10	x		x		
#11		x			x
#12	x				x
#13	x				x
#14		x			x
#15	x				x
#16		x			x
#17	x				x
#18		x			x
#19	x				x
#20		x			x
#21		x		x	
#22	x			x	
#23		x		x	
#24	x			x	
#25		x		x	
#26	x			x	
#27		x		x	
#28	x			x	
#29	x	x		x	
#30		x		x	

Table 8 summarizes characteristics of items in terms of number of solution steps and option format before examining research questions. This table reveals that the final item data has three sets: 10 four option, 10 three options, and 10 NOTA items. Each set has 5 step-solution items and 5 multi-step solutions items.

Item Statistics

Before testing the hypothesis, we report item characteristics of 30 items in terms of item difficulty and item discrimination for three groups, as shown Table 9. Each row shows the items with the same content in different item format.

Table 9. *Item Difficulty and Item Discrimination by Parallel Item Triplets*

Triplet	Item Difficulty (p)			Item Discrimination (r)		
	Four Option	NOTA Option	Three Option	Four Option	NOTA option	Three Option
# 1 ¹	0.70	0.64	0.66	0.45	0.53	0.50
# 2 ¹	0.31	0.29*	0.45	0.47	0.35	0.37
# 3	0.62	0.52	0.52	0.56	0.58	0.43
# 4 ¹	0.64	0.39*	0.58	0.44	0.52	0.49
# 5	0.45	0.48	0.43	0.49	0.53	0.44
# 6	0.37	0.31	0.48	0.43	0.36	0.49
# 7 ¹	0.70	0.54	0.49	0.49	0.37	0.40
# 8	0.66	0.75	0.56	0.44	0.45	0.47
# 9	0.40	0.47*	0.45	0.39	0.31	0.43
# 10 ¹	0.37	0.37	0.47	0.43	0.44	0.44

*NOTA as key; ¹ one-step solution items

The first group includes 10 items with four options, which are conventional items. The item difficulty index (p) of items in the group ranges from 0.31 (# 2) to 0.70 (# 1). This shows

that the items with four options are moderately difficult and easy. The item discrimination index (r) of the items with four options ranges from 0.39 (# 9) to 0.56 (# 3). Since each discrimination index for the items are greater than 0.25, we use all items in the final model.

The second group consists of 10 items with NOTA options where the weakest distractor of conventional items was replaced by NOTA (this process has already reported Chapter 3). The item difficulty index ranges from 0.29 (# 2) to 0.75 (# 8). This means that this group has moderately easy and difficulty items. The item discrimination index of the items with NOTA options are between 0.31 (# 9) and 0.58 (# 3). This makes all items with NOTA options acceptable to use in the final model.

The third group is the items with three options, where the weakest distractor of the conventional items was eliminated (this process has already reported Chapter 3). The item difficulty index of items are moderately easy and difficult, ranging from 0.43 (# 5) to 0.66 (# 1). # 4 has a one-step solution item as well as the easiest one. The item discrimination index of each item with three options are acceptable, ranging from 0.37 (# 2) to 0.50 (# 1).

In summary, the items in each group items has similar item characteristics. Their item difficulty and discrimination indexes generally meet the requirements. Moreover, item solution steps and item difficulty are somewhat associated in each group. The easiest items among the ones with four options and NOTA options are the items with a one-step solution. Table 10 also reports mean of item characteristics for each group regarding to item solution steps (one-step solution and multi-step solutions items).

Table 10. *Mean of Item Characteristic (Difficulty and Discrimination)*

	Item Difficulty			Item Discrimination		
	Four Option	Three option	NOTA Option	Four Option	Three option	NOTA Option
One-step	0.54	0.53	.45	.46	.44	.44
Multi-steps	0.50	0.49	.51	.46	.45	.45
Total	0.50	0.51	.48	.46	.45	.45

Testing Research Questions

Item Difficulty.

For the research question 1 and 3, we examine how item difficulty changes for three types of items (items with four options, three options, and NOTA options) regarding to their solution steps. Item difficulty is used as dependent variable while item solution step, item types, and their interaction are used as independent variables in the multiple regression model, as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \quad (6)$$

y = Item difficulty

x_1 = Item type (N options: items with 4 options, 3 options, and NOTA option)

x_2 = Item solution steps (M steps: one-step vs. multi-steps)

$x_1 x_2$ = Interaction term (item types x item solution steps)

Table 11. *Multiple Regression Model for Item Difficulty*

	Model 1		Model-2	
	Coefficient	Std. Error	Coefficient	Std. Error
constant	0.52*	0.05	0.50*	0.06
M steps	0.01	0.05	0.04	0.09
N options	-0.01	0.06	-0.01	0.09
NOTA option	-0.01	0.06	0.01	0.09
N options x M steps			0.00	0.12
NOTA option x M steps			-0.10	0.12
R-Squared	0.02		0.06	

* $p < 0.05$

Model-1 in Table-11 consists of item solution step and item type as independent variables. The findings shows that item solution step and item types do not influence significantly item difficulty.

Similar results are found in Model-2 which include all independent variables in Model-1 and the interaction terms between item solution steps and item type. Item solution step has no significant impact on item difficulty. Item difficulty mean for four-option MCIs with multi-step solutions is .50. For four-option MCIs with a one-step solution, item difficulty increases by approximately 0.04, but is not significant. Moreover, item difficulty mean for MCIs with four options is not statistically different from MCIs with three options and NOTA option. Moreover, interaction terms are not significant in Model-2. This means that item difficulty is not changed significantly when different item types are applied for MCIs with a one-step solution and multi-step solutions.

These findings reject to null hypothesis for research question-1. Item difficulty decrease statistically when NOTA as an option is applied for MCIs with a one-step solution and multi-step solutions. However, the findings fails to reject null hypothesis for research question-3, which is that item difficulty (p) should not change significantly when the number of options (4 vs. 3) decreases for MCIs with a one-step solution and multi-step solutions.

Item Discrimination.

For the research question 1 and 3, we investigate whether item discrimination changes for MCIs with four options, three options, and NOTA options regarding to item solution steps. Dependent variable is item discrimination while independent variables are item solution steps, item types, and their interaction in the multiple regression model, as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \quad (7)$$

y = Item discrimination

x_1 = Item type (N options: items with 4 options, 3 options, and NOTA option)

x_2 = Item solution steps (M steps: one-step vs. multi-steps)

$x_1 x_2$ = Interaction term (item types x item solution steps)

This model is shown as model-2 in Table-12.

Table 12. *Multiple Regression Model for Item Discrimination*

	Model 1		Model-2	
	Coefficient	Std. Error	Coefficient	Std. Error
constant	0.46*	0.02	0.46*	0.03
M steps	-0.01	0.02	0.00	0.04
N options	-0.01	0.03	-0.01	0.04
NOTA option	-0.01	0.03	-0.01	0.04
N options x M steps			-0.01	0.06
NOTA option x M steps			0.00	0.06
R-Squared	0.01		0.02	

*p<0.05

Model-1 is the same as Model-2, but interaction term between item type and item solution steps. The findings shows that item solution step do not have significant effect on item discrimination. This reveals that item discrimination do not change significantly for MCIs with a one-step solution and multi-step solutions. Item type also does not have a significant effect on item difficulty. This shows that item discrimination are statically the same for MCIs with four options, three options, and NOTA options.

We found the similar results in Model-2 in terms of the impact of item solution step and item type on item discrimination. Interaction terms do not have significant impact on item discrimination in Model-2. This shows that item discrimination is not changed significantly when different item types are applied for MCIs with a one-step solution and multi-steps solutions.

These findings fail to reject null hypothesizes for research question-1 and 3, which is item discrimination do not change significantly when three options or NOTA option is applied for MCIs with a one-step solution and multi-step solutions.

Test Reliability.

For the research question 2 and 4, we calculated the reliability coefficients (Cronbach's alpha estimation methos) for one-step solution and multi-step solutions MCIs with four options, three options, and NOTA options, as shown at Table 13.

We also calculate standard error of estimate (SEE) for each reliability coefficient value to examine whether the reliability coefficient is statistically different from each other. SEE is estimated as (van Zyl, Neudecker, & Nel, 2000, Duhachek & Iacobucci, 2004):

$$SEE = \sqrt{\frac{n}{Q}} \quad (8)$$

where n is sample size and:

$$Q = \left[\frac{2p^2}{(p-1)^2 (\sum_{i=1}^p \sum_{j=1}^p v_{ij})^3} \right] \left[\left(\sum_{i=1}^p \sum_{j=1}^p v_{ij} \right) \left(\sum_{i=1}^p \sum_{k=1}^p v_{ik} v_{ki} + \left(\sum_{i=1}^p v_{ii} \right)^2 \right) - 2 \left(\sum_{i=1}^p v_{ij} \right) \left(\sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^p v_{ik} v_{kj} \right) \right]$$

where p is the number of item and v is covariance matrix among the items. To illustrate, when SEE is calculated for 10 MCIs with four options, three options, and NOTA options in current

study, p is 10. ν is 10 x 10 covariance matrix among each type of MCIs. We used SPSS programming codes to calculate SEE for each item type reliability, as described by Duhachek and Iacobucci (2004).

The findings shows that reliability coefficients (Cronbach's Alpha) are 0.71, 0.70, and 0.67 for MCIs with four options, three options, and NOTA option respectively. Their SEE values are the same to two decimal places: ± 0.02 . This means that the reliability coefficient of the tests with four options and NOTA option are not statistically different because their .95 confidence intervals overlap, (.69, .73) for MCIs with four options and (.65, .70) for MCIs with NOTA option.

Reliability coefficients are 0.77 and 0.78 for MCIs with a one-step solution and multi-step solutions respectively. Their SEE values are the same to two decimal places: ± 0.02 . This means that the reliability coefficient of the tests consisting of MCIs with a one-step solution and multi-step solutions are not statistically different because their .95 confidence intervals overlap, (.75, .79) for MCIs with a one-step solutions and (.76, .80) for MCIs with multi-step solutions.

For MCIs with a one-step solution, reliability coefficients are 0.57 and 0.48 for MCIs with four options and NOTA options respectively. SEE values for MCIs with four options and NOTA options to two decimal places are ± 0.03 and ± 0.04 respectively. This shows that reliability coefficients of the tests consisting of one-step solution MCIs with four options and NOTA options are statistically different because their .95 confidence intervals do not overlap, (.54, .60) for MCIs with four options and (.44, .52) for MCIs with NOTA options.

For MCIs with multi-step solutions, reliability coefficients are 0.58 and 0.54 for MCIs with four options and NOTA options respectively. SEE values for MCIs with four options and NOTA options to two decimal places are the same: ± 0.03 . This shows that reliability coefficients

of the tests consisting of multi-step solutions MCIs with four options and NOTA options are not statistically different because their .95 confidence intervals do not overlap, (.55, .61) for MCIs with four options and (.51, .57) for MCIs with NOTA options.

This findings shows that we reject null hypothesis for research question-2 for MCIs with a one-step solution and fail to reject null hypothesis for research question-2 for MCIs with multi-step solutions.

Table 13. *Test Reliability for Different Set of the Test*

Cronbach's Alpha (α)			
	# of items	Coefficient Alpha (α)	Standard Error with 95%
Four Options	10	0.71	± 0.02
Three Options	10	0.70	± 0.02
NOTA Options	10	0.67	± 0.02
One-Step	15	0.77	± 0.02
Multi-Steps	15	0.78	± 0.02
Four Options – One Step	5	0.57	± 0.03
Four Options – Multi Steps	5	0.58	± 0.03
Three Options – One Step	5	0.53	± 0.04
Three Options – Multi Steps	5	0.57	± 0.03
NOTA Options – One Step	5	0.48	± 0.04
NOTA Options – Multi Steps	5	0.54	± 0.03
All Test	30	0.86	± 0.01

To investigate research question-4, we compares the coefficient alpha for one-step solution MCIs with four options and three options, and multi-step solution MCIs with four options and three options. For MCIs with a one-step solution, reliability coefficients are 0.58 and 0.53 for MCIs with four options and three options respectively. SEE values for MCIs with four options and three options to two decimal places are ± 0.03 and ± 0.04 respectively. This shows

that reliability coefficients of the tests consisting of one-step solution MCIs with four options and three options are not statistically different because their .95 confidence intervals overlap, (.55, .61) for MCIs with four options and (.49, .57) for MCIs with three options.

For MCIs with multi-step solutions, reliability coefficients are 0.58 and 0.57 for MCIs with four options and three options respectively. SEE values for MCIs with four options and three options to two decimal places are the same: ± 0.03 . This shows that reliability coefficients of the tests consisting of multi-step solutions MCIs with four options and three options are not statistically different because their .95 confidence intervals do not overlap, (.55, .61) for MCIs with four options and (.54, .60) for MCIs with three options.

This findings shows that we fail to reject null hypothesis for research question-2 for MCIs with a one-step solution and multi-step solutions.

We summarize that test reliability for the test with one-step solution MCIs with four options is not statistically different than MCIs three options, but MCIs with NOTA options. For MCIs with multi-step solutions, test reliability is not statistically different across MCIs with four options, three options, and NOTA options.

CHAPTER 5

Discussion

Summary of the Study

The purpose of the current study is to empirically examine the impact of conventional MCIs with four options, as compared with MCIs with three options, and MCIs with NOTA options on item characteristics (item difficulty and discrimination) and test characteristic (test reliability). We primarily extend the existing literature by examining the impact of such MCI types on item and test characteristics in terms of item solution steps. The findings show that MCI type and item solution step do not have significant impact on item difficulty and item discrimination and test reliability. The interaction between item type and item solution step also does not significantly influence item difficulty and discrimination. For MCIs with a one-step solution, test reliability does not statistically change across MCIs with four options and three options. However, MCIs with four options are more reliable than NOTA options. For MCIs with multi-step solutions, test reliability does not statistically change across MCIs with four options, three options, and NOTA options.

Item difficulty does not change significantly across MCIs with four options, three options, and NOTA options. Students' test scores for a test with four options are approximately the same as scores on a test with MCIs with three options and NOTA options. Previous studies showed consistent results for item difficulty between MCIs with four options and three options (Delgado & Prieto, 1998; Abad, Olea & Ponsoda, 2001; Shizuka et al., 2006; Baghei & Amrahi, 2011). However, contradictory results are found for item difficulty between MCIs with four options and NOTA options (Tollefson, 1987; Rich & Johanson, 1990; Frary, 1991; Kolstad & Kolstad, 1991; Crehan et al., 1993).

There are some potential reasons we found different results from previous studies regarding item difficulty between MCIs with four and NOTA options. First, it is suggested that the using of NOTA is particularly appropriate for easy conventional MCIs (particularly, $p > 0.60$) (Frary, 1991). In this regard, easy conventional MCIs were generally converted to MCIs with NOTA options in previous studies, which resulted in significantly decreasing item difficulty (Tollefson, 1987; Frary, 1991; Kolstad & Kolstad, 1991; Crehan et al., 1993). However, the conventional MCIs with four options in the current study are not easy ($\mu_{\text{item difficulty}} = 0.52$, (0.31, 0.70)). This causes no change in item difficulty between MCIs with four options and NOTA options. Therefore, item difficulty across MCIs with four options and NOTA options can show different characteristics according to their original level of item difficulty.

Another potential reason for the different in result between previous and current study is that 3 out 10 conventional MCIs randomly chosen to convert to MCIs with NOTA in which the key is replaced “None of the Above” are MCIs with a one-step solution and multi-step solutions. Mean of their item difficulty ($\mu_{\text{item difficulty}} = 0.45$: $p \text{ value}_{\text{item-2}} = 0.31$, $p \text{ value}_{\text{item-4}} = 0.64$, and $p \text{ value}_{\text{item-9}} = 0.40$) is less than mean of all items ($\mu_{\text{item difficulty}} = 0.48$). Previous studies suggested for MCIs with higher p value (easy items), the MCIs with NOTA as key are more difficult than those with NOTA as distractor (Tollefson, 1987; Rich & Johanson, 1990; Frary, 1991). If we select some of such 3 items with higher p values, MCIs rather than MCIs with lower p values, this would cause the item difficulty of MCIs with NOTA to decrease, according to existing literature. As a result, we conclude that MCIs with NOTA options may be more difficult than conventional MCIs.

We also found that item solution step does not have significant impact on item difficulty. This is surprising result because it is expected that MCIs with a one-step solution are easier than

MCIs with multi-step solutions. Some of MCIs with a one-step solution is long question. The length of item can be a potential reason to make the items more difficulty.

Item discrimination does not change significantly across conventional MCIs with four options, three options, and NOTA options. This is consistent with previous studies (Delgado&Prieto, 1998; Frary, 1991; Knowles&Welch, 1992; Crehan et al., 1993; DiBattisa, Sinnige-Egger & Foruna, 2013). This shows that MCIs with four options, three options, and NOTA options have statistically similar characteristics in discriminating the extent to which high ability students answer correctly while low ability students answer incorrectly. However, the range of item discrimination for MCIs with NOTA options is slightly greater than those for MCIs with four and three options ($\text{range}_{\text{four-option}} = (0.39, 0.56)$; $\text{range}_{\text{three-option}} = (0.37, 0.50)$; $\text{range}_{\text{NOTA-option}} = (0.31, 0.58)$). The items with lowest discrimination index among the MCIs with NOTA options are those with NOTA as key, which makes the range of discrimination index increase. The recent study showed that the item discrimination index for MCIs with NOTA as key is smaller than those with distractor (DiBattisa, Sinnige-Egger & Foruna, 2013)

Additionally, our findings indicate that item characteristics do not change significantly across MCIs with four options, three options, and NOTA options in terms of item solution step. This shows that item difficulty does not change across one-step solution MCIs with four options, three options, and NOTA options, as well as multi-step solutions MCIs with four options, three options, and NOTA options. Table 11 shows that mean of multi-step solutions MCIs with four, three, and NOTA options are similar ($\mu_{\text{item difficulty - four}} = .50$; $\mu_{\text{item difficulty - three}} = .49$; $\mu_{\text{item difficulty - NOTA}} = .51$), although one-step solution MCIs with four options are similar to MCIs with three options and greater than MCIs with NOTA options ($\mu_{\text{item difficulty - four}} = .54$; $\mu_{\text{item difficulty - three}} = .53$; $\mu_{\text{item difficulty - NOTA}} = .45$). However, we found insignificant results in regression model. These

results are surprising results because recent study addressed by Rodriguez (2005) showed that MCIs with three options are significantly more difficult than MCIs with four options ($p_{\text{three-options}} < p_{\text{four-options}}$) although the difference of p-values is 0.04.

There are some potential reasons why we found insignificant results. First, the sample size of the regression models is 30, which is total number of MCIs applied in the current study. Sample size is an important factor to obtain statistical power. A model with an inadequate sample is more likely to cause insignificant results. Traditionally, power analysis is used to efficiently interpret insignificant results. However, power analysis was not eligible tool for the regression model of the current study because each item's difficulty index was derived from the responses of 1667 students, not a single student.

In addition, item discrimination as an item characteristic does not change across one-step solution MCIs with four options, three options, and NOTA options as well as multi-step solutions MCIs with four options, three options, and NOTA options. The mean of item discrimination for MCIs with a one-step solution are approximately the same ($\mu_{\text{item discrimination - four}} = .46$; $\mu_{\text{item discrimination - three}} = .44$; $\mu_{\text{item difficulty - NOTA}} = .44$), as shown in Table 12. The mean of item discrimination for MCIs with multi-step solutions are also approximately the same ($\mu_{\text{item discrimination - four}} = .46$; $\mu_{\text{item discrimination - three}} = .45$; $\mu_{\text{item difficulty - NOTA}} = .45$). As a result, item discrimination does not change across different type of MCIs in terms of item solution steps.

For MCIs with a one-step solution, the findings showed that test reliability for MCIs with four options is not statistically different from MCIs with three options. This is consistent with previous studies (Trevison, 1991; Delgado & Prieto, 1998). However, MCIs with four options is statistically different from MCIs with NOTA options. This is contradictory result with previous studies, but one study by Tollefson (1987). For MCIs with multi-step solutions, test reliability

for MCIs with four options is not statistically different from MCIs with three options and NOTA options.

Implications for Testing Practice

The findings of this study make important contributions. This study helps item writers and teachers to more quickly and easily construct reliable and valid MCIs. Using MCIs with three options is less challenging and saves time for item writers because they require a fewer number of plausible distractors than conventional MCIs with four options. This is crucial and difficult part of constructing a MCI.

Moreover, administering a test with MCIs with three options can increase test reliability in three ways, as compared with conventional MCIs with four options. First, MCIs with three options are administered in less time than conventional MCIs with four options. Shorter test decreases students' fatigue and test anxiety, which causes test reliability to increase. Second, if a test with MCIs with three options is applied at the same administration time as a test with MCIs with four options, more items are added to the test. The test with more items is more likely to be more reliable. Third, when a test with MCIs with three options is applied at the same administration time as a test with MCIs with four options, students are likely to have more time to read all individual questions in the test, which decreases their test anxiety about time limit. This leads to increases test reliability.

Limitations and Directions for Future Research

There are several limitations of this study. First, the test administered in the study included only the math items constructed based on particular a math content area according to 7th grade curriculum in Turkey and the U.S.A. Future research needs to examine how psychometric

characteristics of each MCI type are changed when the math tests with different content areas or the tests measuring different subjects than math are administered to students from the same grade level and different grades from randomly chosen schools. Therefore, we would be able to make generalizations from our current findings and apply them to the larger population.

Second, the current study included 30 items: 10 MCIs with four options, three options, and NOTA options. The regression models in the current study were not statistically significant because there are only 10 MCIs for each MCI type. In future studies, at least 30 MCIs for each MCI type should be applied to obtain more accurate results.

Third, the test constructed in the current study includes MCIs with four options, NOTA options, and three options, respectively. This test was administered in this fixed order to all of the students. This causes a systematic error due to order effect. In the future, the tests should be designed to counterbalance the order of MCIs in order to decrease error.

Fourth, when converting conventional MCIs with four options to MCIs with NOTA options and three options, we selected an option to be eliminated by applying a non-parametric IRT model. In the future, the options should be eliminated by different methods from the current study so that researchers can examine whether item and test characteristics remain the same.

Fifth, we applied the bi-factor IRT model to determine item solution step. One of the limitations of this model as currently implemented is the orthogonality assumption, in which there is no correlation between domain and subdomains, and within subdomains. However, the test administered in the current study includes only math items which are likely to correlate with each other because they were constructed under the same domain (Equation and Expression (EE) in CCSS), but different content areas (EE1, EE2, EE3, and EE4). In future experiments, researchers need to apply different methods, such a confirmatory factor analysis and hierarchical

linear modeling, to determine whether the item solution step characteristics of each item remain the same.

References

- Abad, F., Olea, J., & Ponsoda, V. (2001). Analysis of the optimum number alternatives from the Item Response Theory. *Psicothema*, 13(1), 152-158.
- Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175-185.
- American Psychological Association, American Educational Research Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, 53(2), 192-211.
- Bormuth, J.R. (1970). *On a theory of achievement test items*. Chicago: University of Chicago Press.
- Burton, S.J., Sudweeks, R. R., Merrill, P. F. & Wood, B. (1990). How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty. *Brigham Young University.Dept. of Instructional Science*. Retrieved from <http://testing.byu.edu/info/handbooks/betteritems.pdf>
- Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO 2.1 for Windows. *Chicago, IL: Scientific Software International*.
- Case, S. M., & Swanson, D. B. (1998). *Constructing written test questions for the basic and clinical sciences* (2nd ed.). Philadelphia, PA: National Board of Medical Examiners.
- Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions

- for continuing medical education activities and self-assessment modules. *RadioGraphics*, 26(2), 543-551.
- Common Core State Standards Initiative (CCSSI). (2010). Common Core State Standards for Mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98-104.
- Crehan, K. D., & Haladyna, T. M. (1991). The validity of two item-writing rules. *Journal of Experimental Education*, 59(2), 183-192.
- Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, 53(1), 241-247.
- Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, 14(3), 197-201.
- Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education*, 10(2), 133-143. doi:10.1007/s10459-004-4019-5
- Dudycha, A.L., & Carpenter, J.B. (1973). Effects of item format on item discrimination and item difficulty. *Journal of Applied Psychology*, 58 (1), 116-121.
- Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): an accurate and precise confidence interval estimate. *Journal of applied psychology*, 89(5), 792.
- Eubank, R. L. (1988). Spline smoothing and nonparametric regression. New York: Marcel

- Dekker.
- Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education*, 4(2), 115–124.
- Frey, B.B., Petersen, S., Edwards, L.M., Pedrotti, J.T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, 21(4), 357-364.
- Gibbons, R.D., & Hedeker, D.R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3), 423-436.
- Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 37-50.
- Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 51-78.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-334.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park: Sage Publications.
- Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing test banks. *The Journal of Education for Business* 73(2), 94-97.
- Henrysson, S. (1971). Analyzing the test item. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 130-159) Washington, DC: American Council on Education.
- Hughes, H.H. & Trimble, W. E. (1965). The use of complex alternatives in multiple-choice terms. *Educational and Psychological Measurement*, 25(1), 117-126.

- Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (2011). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education*, 24(3), 210-234.
- Kingston, N.M., & Kramer, L.M.B. (2013). High-stakes test construction and test use. In T.D. Little (Ed.), *The Oxford Handbook of Quantitative Methods: Statistical Analysis* (pp. 189-205). Oxford University Press.
- Knowles, S. L. and Welch, C. A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using "none-of-the-above". *Educational and Psychological Measurement*, 52(3), 571-577.
- Kolstad, R. K., & Kolstad, R. A. (1991). The option "none of these" improves multiple-choice test items. *Journal of Dental Education*, 55(2), 161–163.
- Landrum, R. E., Cashin, J. R., & Theis, K. S. (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement*, 53(3), 771–778.
- Lee, Y. (2007). A Comparison of Methods for Nonparametric Estimation of Item Characteristic Curves for Binary Items. *Applied Psychological Measurement*, 31(2), 121–134.
- Lord, F. M. (1953). An examination of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 18(1), 57-76.
- McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher*, 26(8), 709-712.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.
- Owen, S. V., & Froman, R. D. (1987). What's wrong with three-option multiple choice items? *Educational and psychological measurement*, 47(2), 513-522.

- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611-630.
- Reise, S. P., Morizot, J., & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16, 19–31.
- Rich, C. E., & Johanson, G. A. (1990, April). *An item-level analysis of “none of the above.”* Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3-13.
- StataCorp, L. P. (2013). Stata User’s Guide. *College Station, TX: Stata Press, StataCorp LP.*
- Shizuka, T., Takeuchi, O., Yashima, T. & Yoshizawa, K. (2006). A comparison of three-and four-option English tests for university entrance selection purposes in Japan. *Language Testing*, 23(1), 35-57.
- Simonoff, J. S. (1996). Smoothing methods in statistics. New York: Springer.
- Statistics, I. I. S. (2012). Core system user’s guide. *Chicago: SPSS Inc.*
- Thorndike, R. M. (2005). Measurement and evaluation in psychology and education (Seventh Edition). Upper Saddle River, NJ: Pearson Education.
- Tollefson, N. (1987). A comparison of the item difficulty and item discrimination of multiple-choice items using the " none of the above" and one correct response options. *Educational and psychological measurement*, 47(2), 377-383.
- Trevisan, M. S., Sax, G., & Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement*, 51, 829–837.

- Van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, 65, 271–280.
- Wesman, A.G., & Bennett, G. K. (1946). The use of “none of these” as an option in test construction. *Journal of Educational Psychology*, 37(9), 541-549.

Appendix A
Multiple – Choice Items

Q1. Which expression could be used to find 3 times 5 more than x ?

- A. $3x + 5$
- B. $5x + 3$
- C. $3(x + 5)$
- D. $(5 + 3)x$

Q2. The cost of a small pizza is x dollars at Green Pizza. Mark has a 20% off coupon. Which expression could be used to find how much he pays for a small pizza at Green Pizza?

- A. $0.20x$
- B. $0.20 + x$
- C. $x - 0.20$
- D. $x - 0.20x$

Q3. The price of a camera is \$300. There is a sales tax of 10%. How much does it cost to buy 2 cameras?

- A. \$660
- B. \$630
- C. \$620
- D. \$600

Q4. Which expression could be used to find 12 more than $\frac{7}{5}$ of x ?

- A. $\frac{7}{5}x + 12$
- B. $x + \frac{7}{5} + 12$
- C. $\frac{7}{5}(x + 12)$
- D. $x(\frac{7}{5} + 12)$

Q5. Jay has \$80. He uses 60% of his money to buy a pair of shoes. Then, he spends $\frac{1}{4}$ of the remaining money to buy a gift for his friend. How much money does he pay for the gift?

- A. \$12
- B. \$10
- C. \$8
- D. \$5

Q6. Martha reads $\frac{3}{5}$ of a book on Monday. She reads $\frac{1}{4}$ of the remaining pages on Tuesday. What percentage of the book does she read on Tuesday?

- A. 10%
- B. 15%
- C. 30%
- D. 35%

Q7. Which expression could be used to find 4 times as many as 6 less than x ?

- A. $(6)(4) - x$
- B. $x - (4)(6)$
- C. $4x - 6$
- D. $4(x - 6)$

Q8. Jim bought 5 packs of pens. He paid \$25 total, which includes \$0.40 in tax per pack of pens. What is the price of a pack of pens before tax?

- A. \$4.60
- B. \$4.92
- C. \$5.08
- D. \$5.40

Q9. The capacity of an elevator is 300 kg. John and 5 friends can use the elevator safely. John weighs 60 kg. Which must be true for the average weight (x) of each of John's 5 friends?

- A. $x > 48$
- B. $x < 48$
- C. $x > 72$
- D. $x < 72$

Q10. Karla reads x pages of a book in 3 hours. Which expression could be used to find the number of pages Karla reads in 2 hours?

- A. $\frac{2x}{3}$
- B. $\frac{3x}{2}$
- C. $\frac{x}{2} + 3$
- D. $\frac{x}{3} + 2$

Q11. The price of a pair of shoes is \$40. The price of a coat is 20% more than the price of a pair of shoes. What is the cost of 3 coats?

- A. \$120
- B. \$128
- C. \$144
- D. None of the above

Q12. Which expression could be used to find 8 times as many as 5 less than x ?

- A. $(5)(8) - x$
- B. $8x - 5$
- C. $8(x - 5)$
- D. None of the above

Q13. The cost of 4 T-shirts is x dollars. Which expression could be used to find the cost of 5 T-shirts?

- A. $\frac{5x}{4}$
- B. $\frac{4x}{5}$
- C. $\frac{x}{4} + 5$
- D. None of the above

Q14. John's rent costs $\frac{3}{8}$ of his salary. He also pays for utilities that cost $\frac{2}{5}$ of the remaining salary after rent is paid. What percentage of his salary do utilities cost John?

- A. 12.5%
- B. 15%
- C. 25%
- D. None of the above

Q15. Which expression could be used to find 8 times 10 more than x ?

- A. $8x + 10$
- B. $10x + 8$
- C. $8(x + 10)$
- D. None of the above

Q16. Melissa spent \$120 on a trip. She paid \$30 to rent a car, and stayed at a hotel for 2 nights. How much did the hotel cost per night?

- A. \$30
- B. \$45
- C. \$75
- D. None of the above

Q17. Which expression could be used to find 10 more than $\frac{1}{4}$ of x ?

- A. $x(\frac{1}{4} + 10)$
- B. $x + \frac{1}{4} + 10$
- C. $\frac{1}{4}(x + 10)$
- D. None of the above

Q18. There are 200 vehicles in the parking lot. 25% of the vehicles are trucks while others are cars. $\frac{2}{5}$ of the cars are white. How many cars are white in the parking lot?

- A. 60
- B. 50
- C. 20
- D. None of the above

Q19. The cost of gas per gallon was x dollars in June. The cost was increased by 6% in July. Which expression could be used to find the cost of gas per gallon in July?

- A. $0.6x$
- B. $0.6 + x$
- C. $x - 0.6x$
- D. None of the above

Q20. Tim has more than 180 math questions to solve for a project. His friend helps him solve 60 of them. Which inequality shows the average number of questions (x) per day Tim needs to solve to complete the project in the next 5 days?

- A. $x < 24$
- B. $x > 48$
- C. $x < 48$
- D. None of the above

Q21. Marry took a test with questions from several subject areas. $\frac{2}{5}$ of the questions were math questions. $\frac{1}{2}$ of the remaining questions were science questions. What percentage of the test were science questions?

- A. 20%
- B. 30%
- C. 40%

Q22. The price of a computer is x dollars. The sales tax rate is 10%. Which expression could be used to find the total cost of the computer including tax?

- A. $0.1x$
- B. $x - 0.1x$
- C. $x + 0.1x$

Q23. Johnson rented a truck from a rental company. The cost of the rental truck was \$20 per day and \$0.80 for each mile driven. His bill was \$60. How many miles did he drive?

- A. 50 miles
- B. 35 miles
- C. 28 miles

Q24. Which expression could be used to find 12 times 16 more than x ?

- A. $12x + 16$
- B. $16x + 12$
- C. $12(x + 16)$

Q25. Jack has \$200 in his savings account. He earns \$10 per hour at work, and deposits all of the money he earns into his savings account. At the end of the month, Jack wants to have more than \$1000 in the account. How many hours will he need to work this month?

- A. $x > 120$
- B. $x > 80$
- C. $x < 80$

Q26. Which expression could be used to find 20 more than $\frac{1}{2}$ of x ?

- A. $\frac{1}{2}x + 20$
- B. $x + \frac{1}{2} + 20$
- C. $\frac{1}{2}(x + 20)$

Q27. The price of a T-shirt is \$15. When two T-shirts are purchased, there is a 20% discount. Amy buys 2 T-shirts. How much does she pay?

- A. \$24
- B. \$26
- C. \$27

Q28. Which expression could be used to find 12 times as many as 16 less than x ?

- A. $(16)(12) - x$
- B. $12x - 16$
- C. $12(x - 16)$

Q29. A half-cup of walnuts contains x grams of protein. Which expression could be used to find the amount of protein in three cups of walnuts?

- A. $\frac{3x}{\frac{1}{2}}$
- B. $\frac{\frac{1}{2}x}{3}$
- C. $\frac{x}{\frac{1}{2}} + 3$

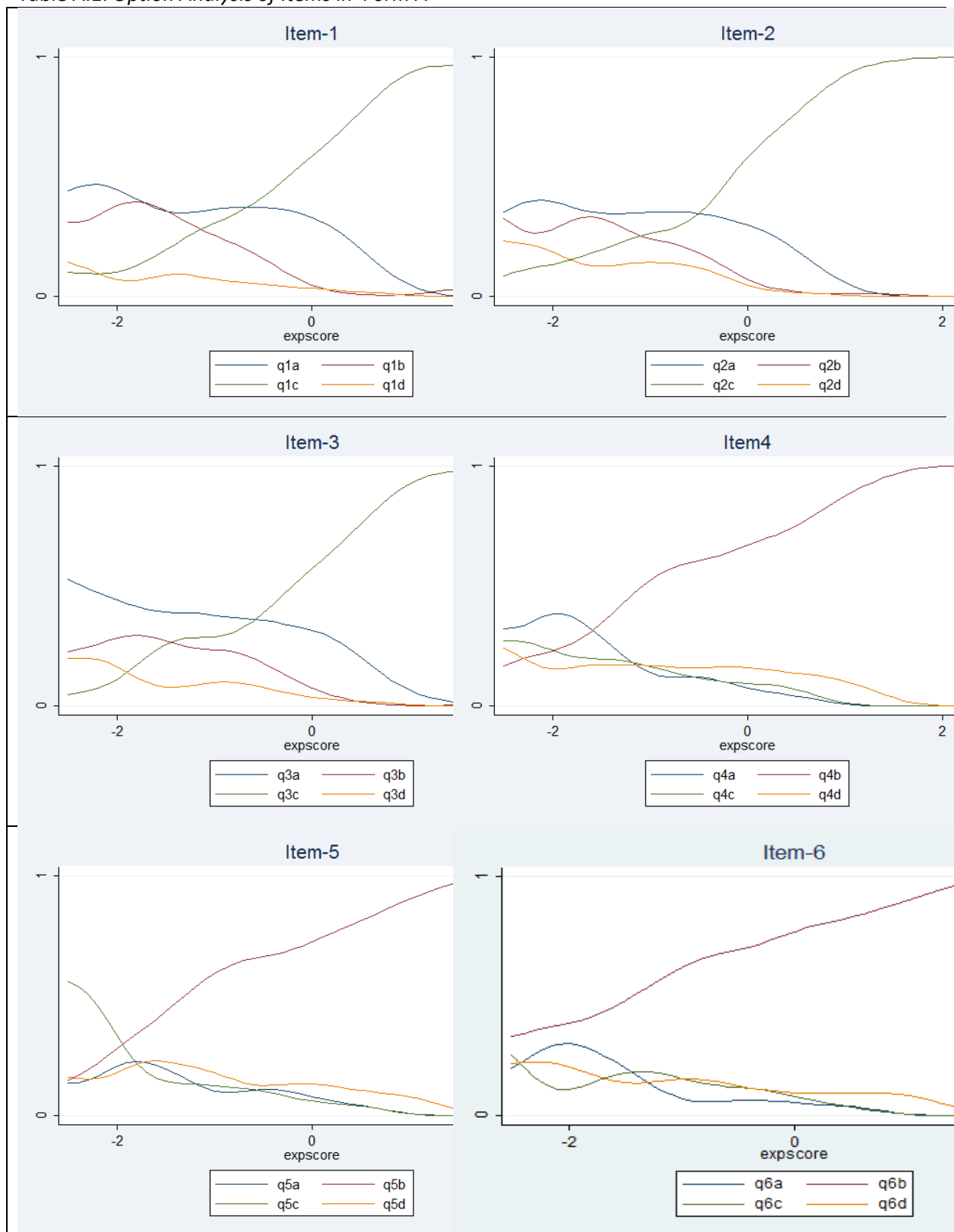
Q30. William had 150 math questions to complete over the weekend. He solved 30% of the questions on Saturday. He solved $\frac{3}{5}$ of the remaining questions on Sunday. How many questions did William finish on Sunday?

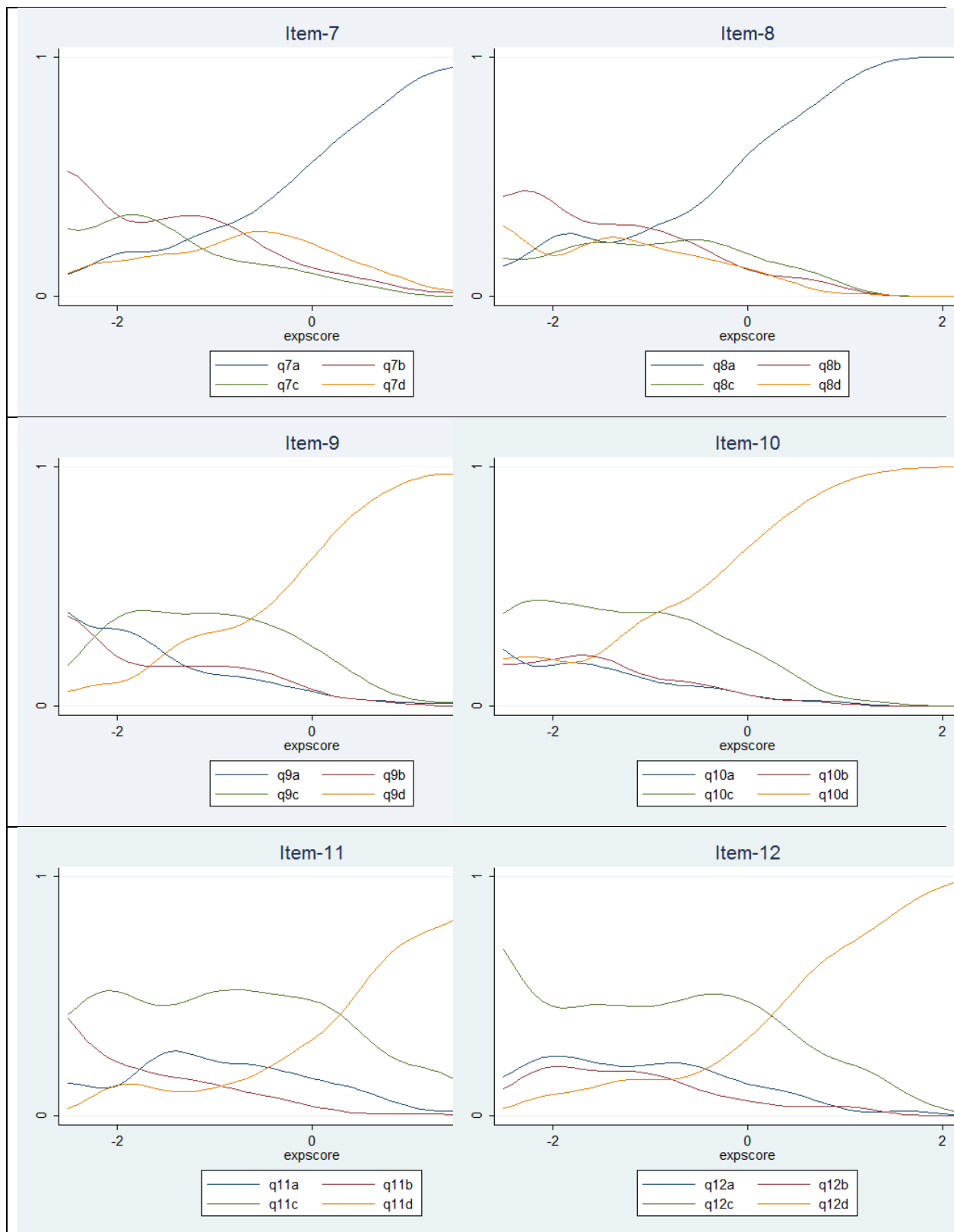
- A. 27
- B. 48
- C. 63

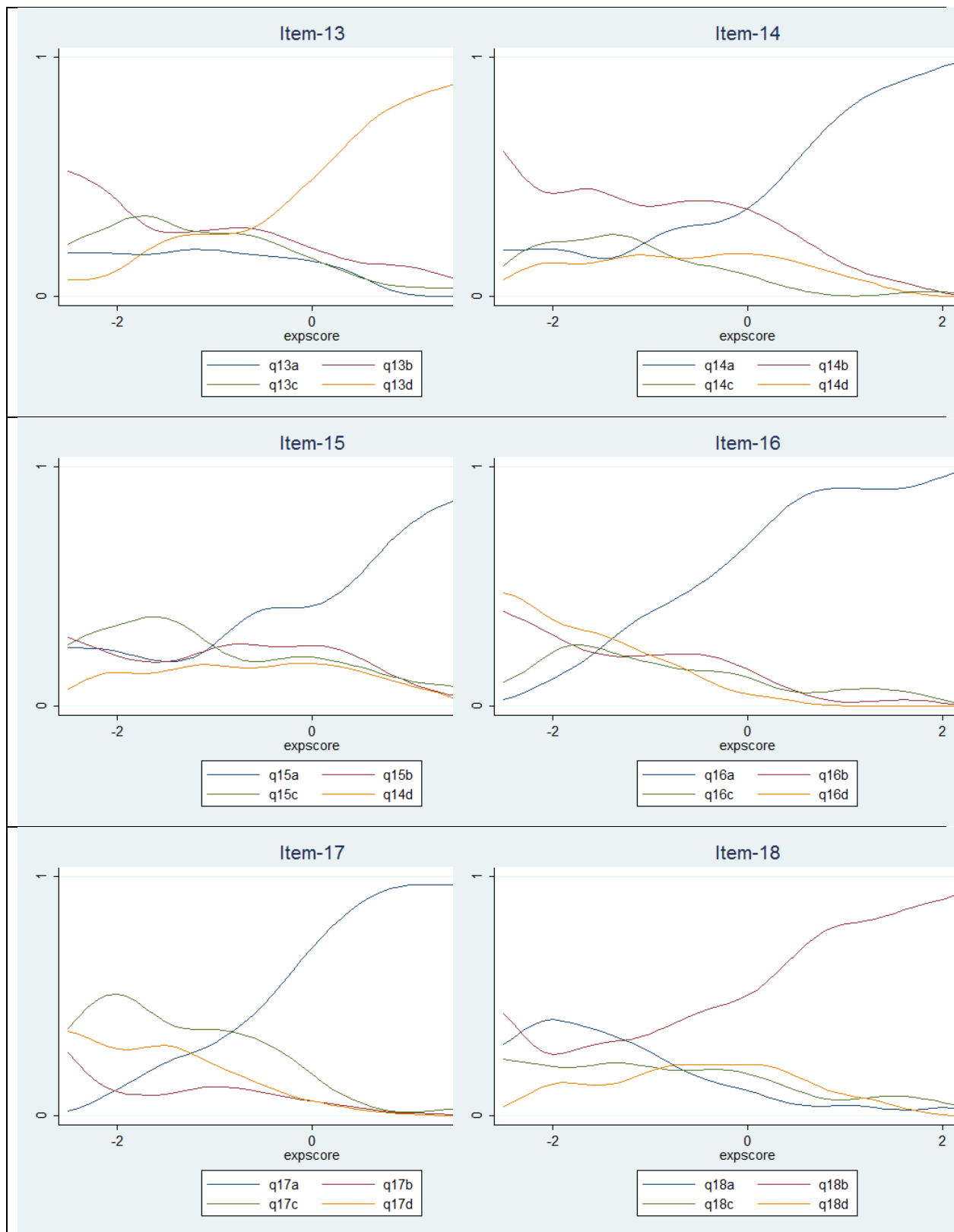
Appendix B

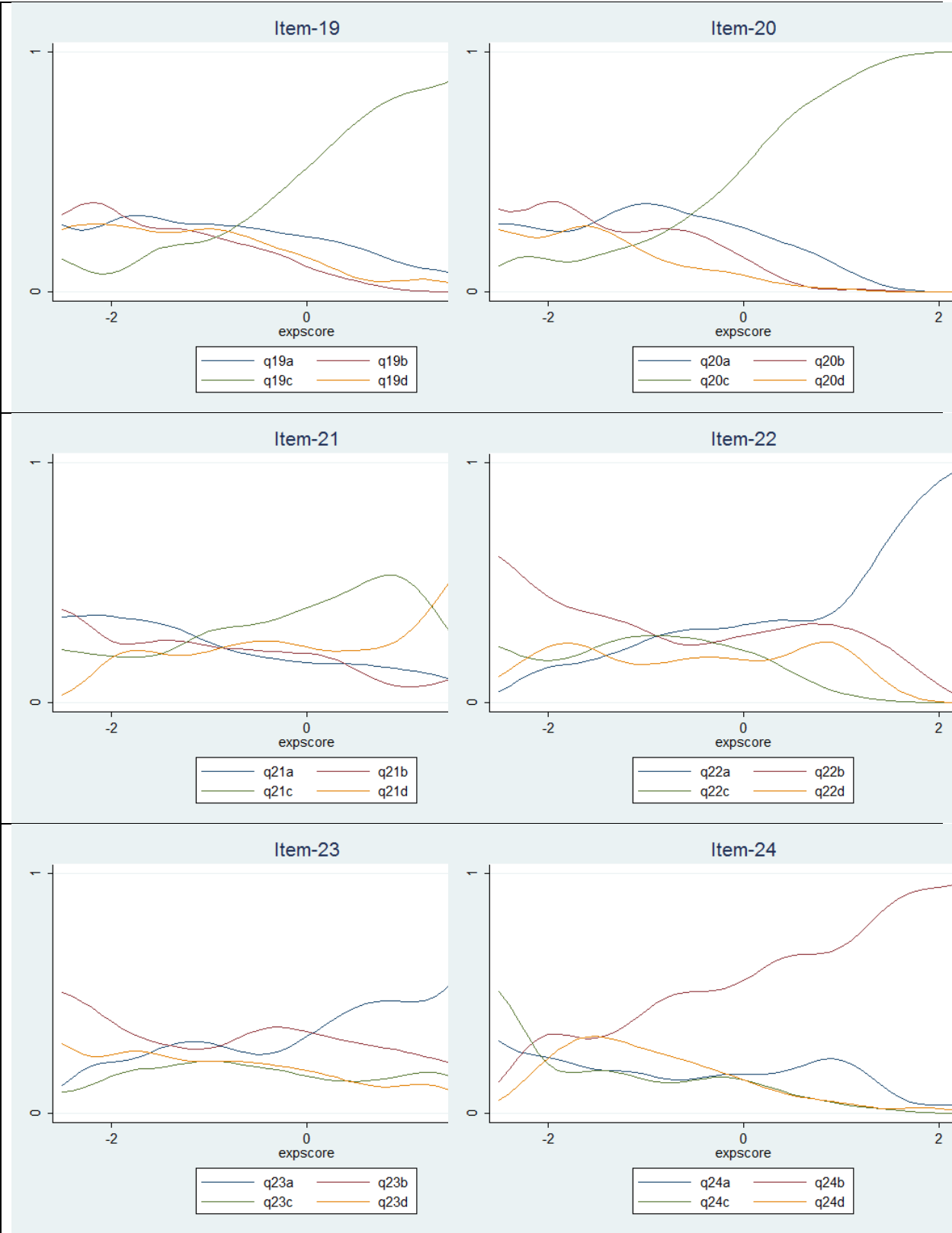
Option Analysis of Items

Table A.1: Option Analysis of Items in Form A









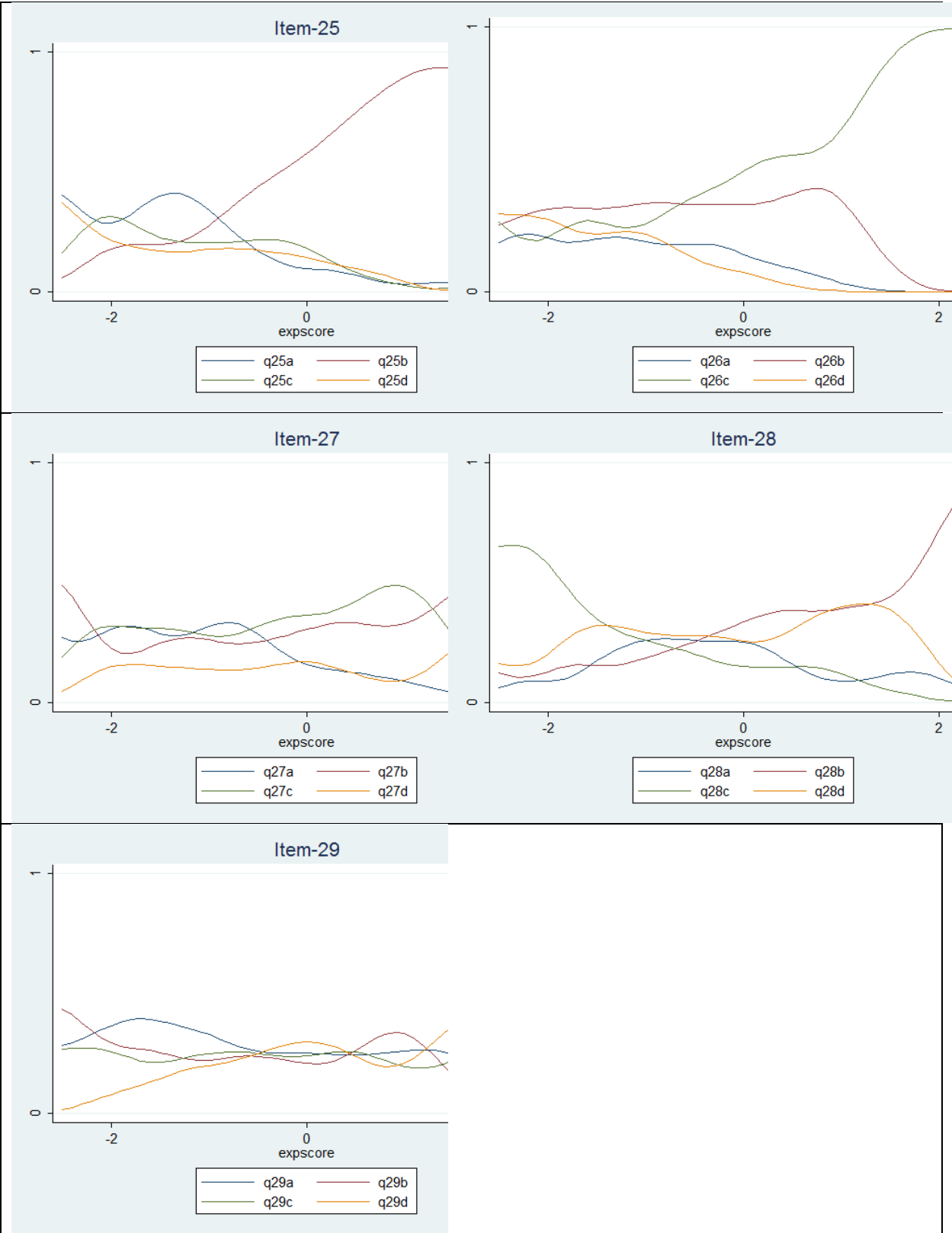
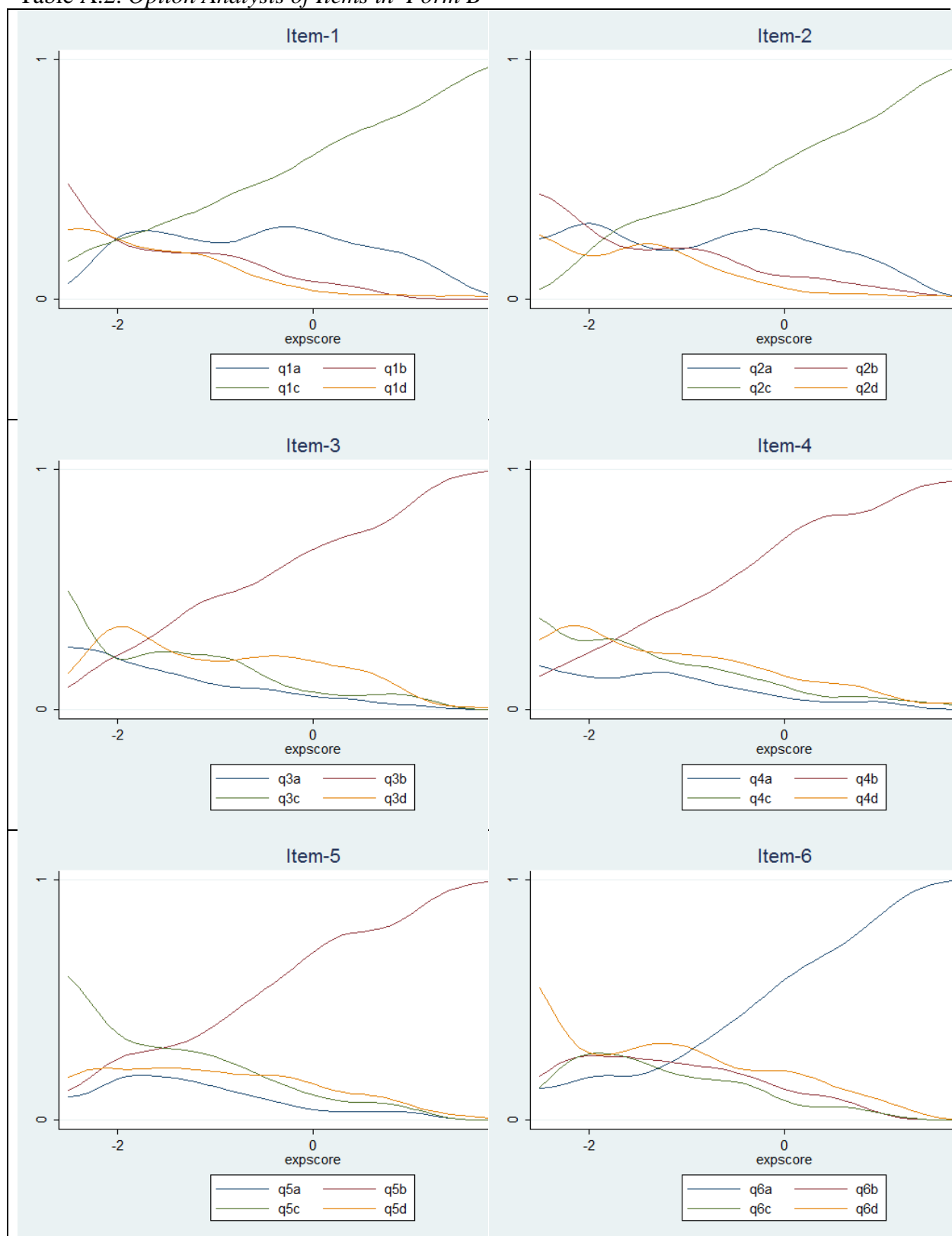
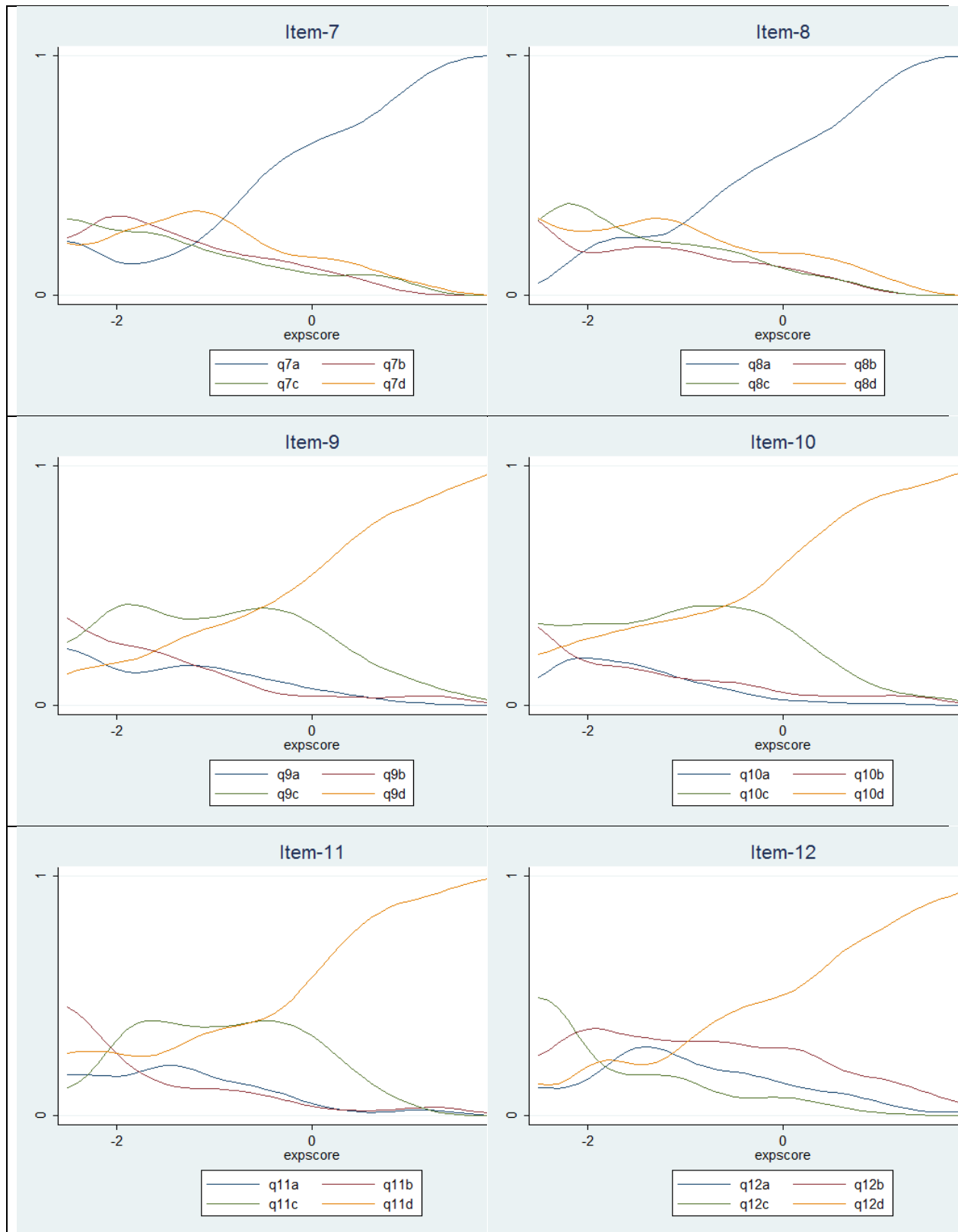
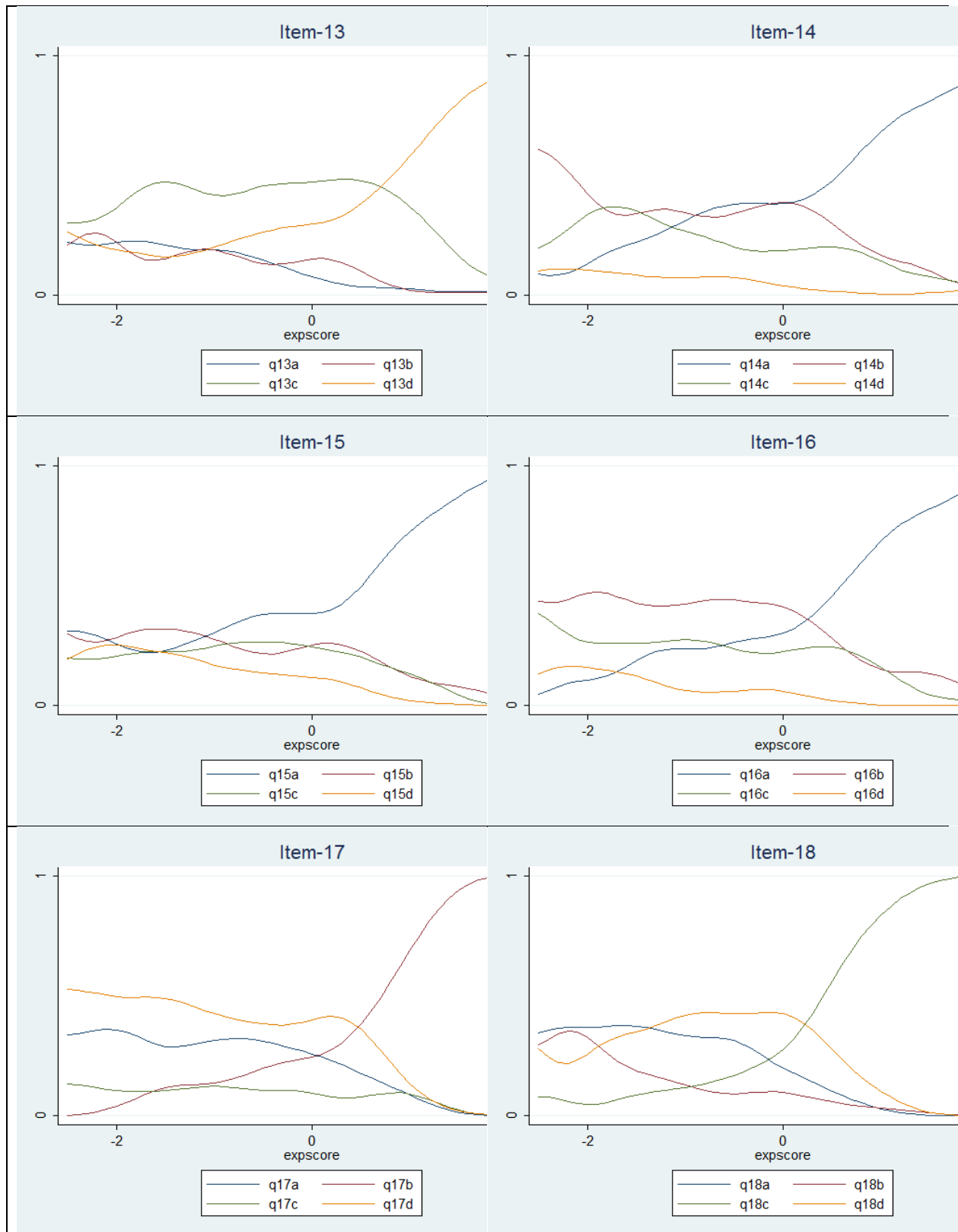
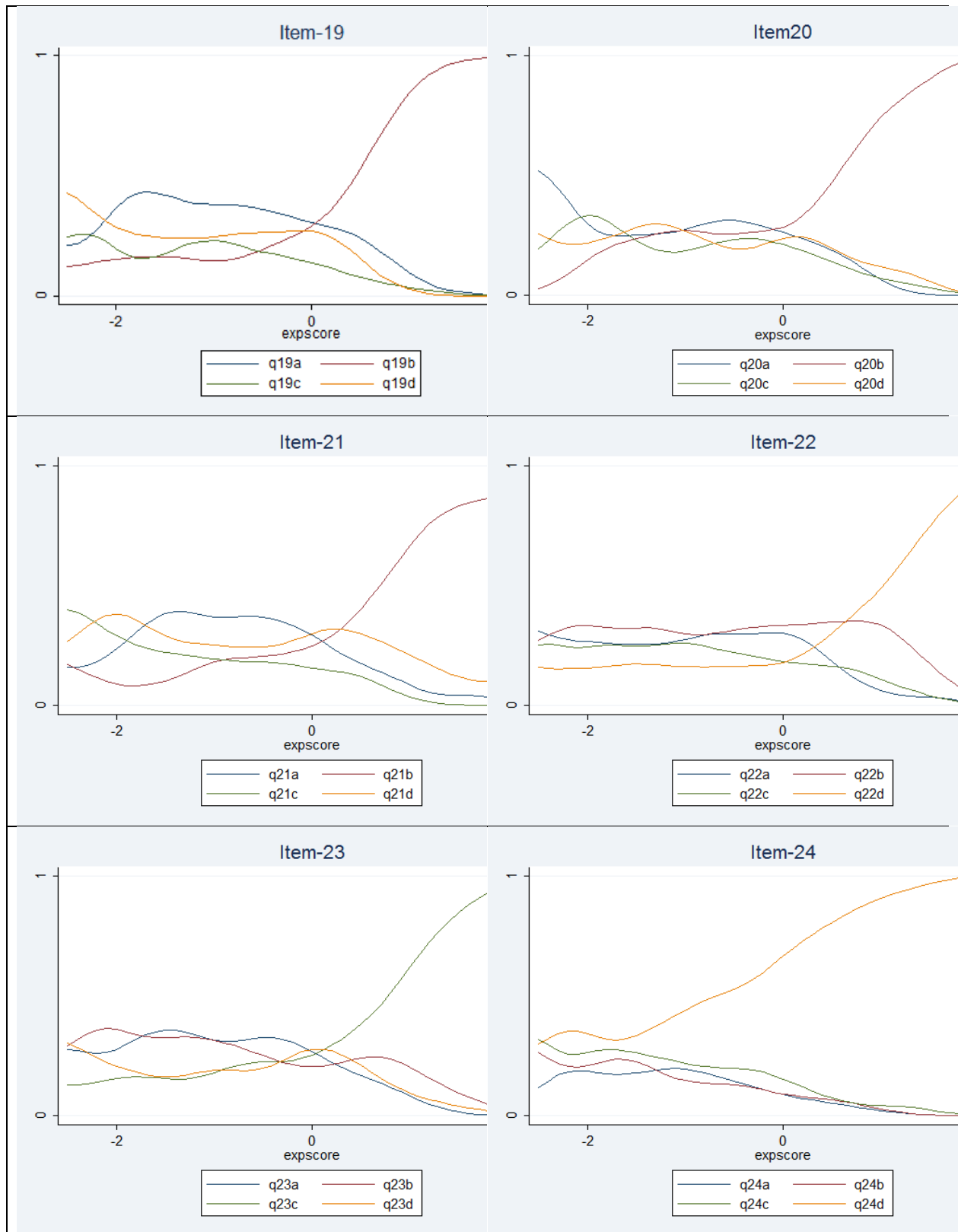


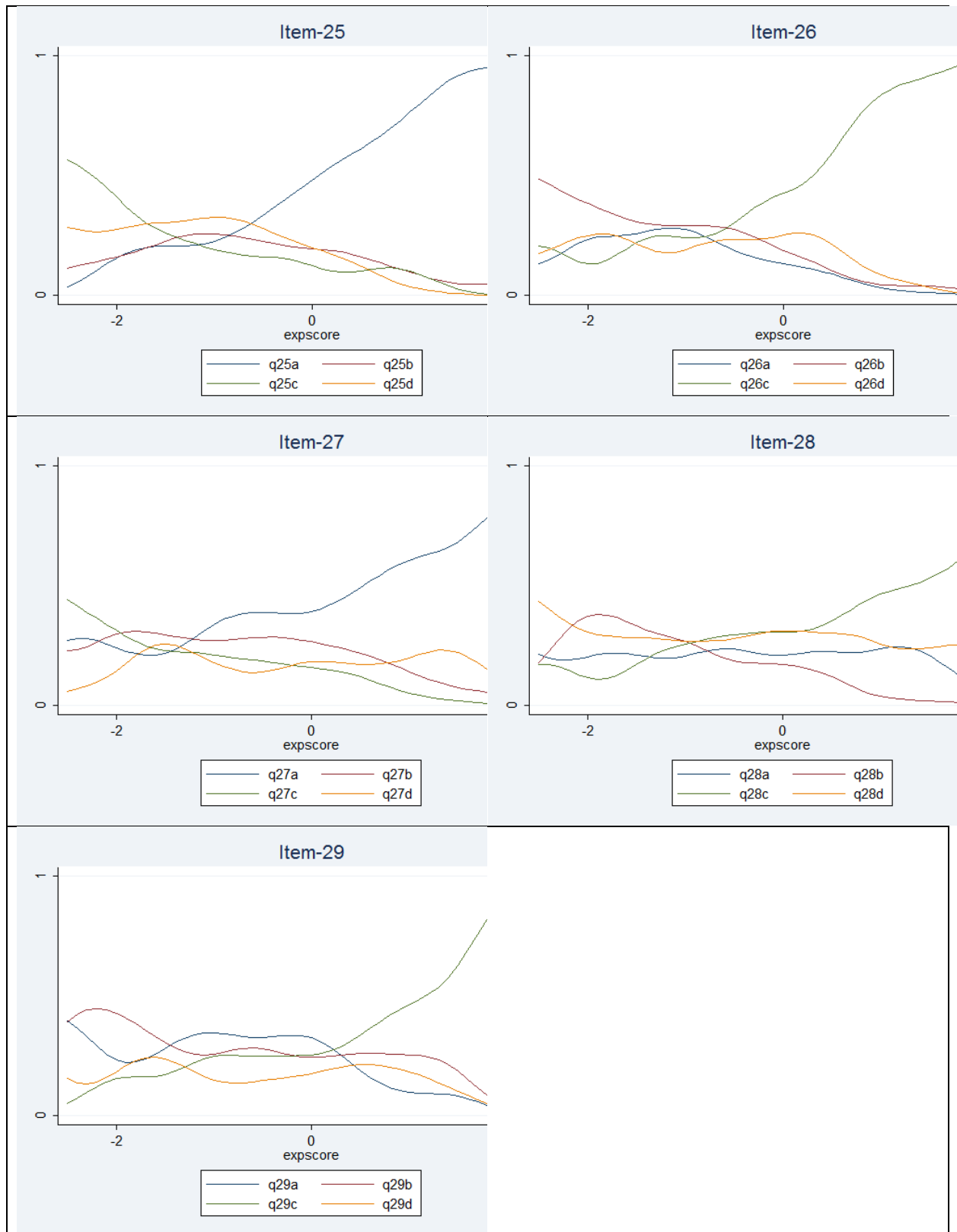
Table A.2: Option Analysis of Items in Form B











Appendix C

Factor Loading for Items in Bi-Factor IRT Model

Factor Loadings Values

Item	General Factor		One-Step		Multi-Step	
	Loading	Standard Error	Loading	Standard Error	Loading	Standard Error
1	0.60	0.06	0.30	0.13	0.00	0.00
2	0.59	0.05	-0.13	0.08	0.00	0.00
3	0.65	0.05	0.00	0.00	0.46	0.07
4	0.62	0.05	-0.12	0.08	0.00	0.00
5	0.51	0.05	0.00	0.00	0.38	0.08
6	0.43	0.05	0.00	0.00	0.32	0.09
7	0.66	0.06	0.38	0.13	0.00	0.00
8	0.48	0.06	0.00	0.00	0.38	0.08
9	0.43	0.05	0.00	0.00	0.06	0.09
10	0.49	0.05	0.05	0.07	0.00	0.00
11	0.64	0.05	0.00	0.00	0.49	0.07
12	0.33	0.08	0.79	0.09	0.00	0.00
13	0.50	0.05	0.08	0.07	0.00	0.00
14	0.35	0.06	0.00	0.00	0.21	0.09
15	0.70	0.05	0.32	0.13	0.00	0.00
16	0.63	0.05	0.00	0.00	0.18	0.08
17	0.66	0.04	-0.13	0.07	0.00	0.00
18	0.56	0.05	0.00	0.00	0.38	0.08
19	0.41	0.05	-0.13	0.07	0.00	0.00
20	0.26	0.05	0.00	0.00	0.13	0.09
21	0.54	0.05	0.00	0.00	0.17	0.08
22	0.42	0.05	-0.11	0.07	0.00	0.00
23	0.57	0.05	0.00	0.00	0.09	0.09
24	0.69	0.05	0.16	0.12	0.00	0.00
25	0.47	0.05	0.00	0.00	0.08	0.09
26	0.63	0.05	-0.07	0.07	0.00	0.00
27	0.46	0.05	0.00	0.00	0.20	0.09
28	0.38	0.06	0.63	0.06	0.00	0.00
29	0.50	0.05	0.09	0.08	0.00	0.00
30	0.47	0.05	0.00	0.00	0.15	0.09